A CORRIDOR-LEVEL PEDESTRIAN CRASH RISK ASSESSMENT FRAMEWORK USING AUTONOMOUS VEHICLE SENSOR DATA

Sunny Singh

MAdvEng(Transport), BE(Civil)

Submitted in fulfilment of the requirements for the degree of

Master of Philosophy

School of Civil and Environmental Engineering

Faculty of Engineering

Queensland University of Technology

2023

Keywords

Autonomous vehicle; extreme value modelling; pedestrian safety analysis; vulnerable road user; vehicle-pedestrian conflict.

Abstract

The road network is a critical element of our infrastructure that facilitates social mobility and supports economic growth. However, the rising number of global crashes on our roads that leads to countless injuries and loss of lives remains a significant concern for governments and transport engineering professionals. To address this pressing issue, road safety analysis has long been a crucial component of infrastructure planning guidelines designed by transportation engineering professionals.

Traditional crash-based safety analysis techniques often employed by transport engineers suffer from limitations such as underreporting, logging errors, and limited behavioural information. Contrasting to crash-based data used for such modelling, a growing interest in utilising traffic conflicts and linking them to crashes has been noticed; however, recent studies have evaluated road user crash risk using traffic conflicts at either single or multiple but scattered intersections. Comparatively, less attention has been paid to corridor-wide safety because of data collection scalability limitations. While these studies have provided valuable insights, they often fail to capture the complexities and interdependencies of crash mechanisms along transportation corridors. As most road user journeys are not limited to a single intersection and comprise multiple road segments forming a corridor, road safety analysis at a corridor level becomes paramount. With multiple recent autonomous vehicle trials on public roads generating massive amounts of road user data, the use of such rich information to resolve data collection scalability issues and perform corridor-wide safety analysis is somewhat limited where it appears to be most relevant.

This thesis proposes an extreme value theory modelling framework to estimate corridor-wide pedestrian crash risk using autonomous vehicle sensor data. In particular, the study estimates two extreme value models, including Block Maxima extreme sampling approach relating to Generalised Extreme Value distribution and Peak Over Threshold extreme sampling approach relating to Generalised Pareto distribution. The autonomous vehicle data for model estimation was obtained from a publicly available source, Argoverse. Their autonomous vehicle fleet, operating in six different cities in the USA, is equipped with two 64 beams synchronised LiDAR sensors, a cluster of seven high-resolution cameras, and a pair of stereo-vison high-resolution cameras to capture

surrounding road users' information. Through a case study of a selected transportation corridor focussing on an arterial corridor in Miami, USA, this thesis assesses the application of the proposed corridor-wide road safety analysis framework. A subset of the Argoverse dataset focusing on the selected corridor was used to extract vehicle trajectories and pedestrians within a range of 200 m from the autonomous vehicles that travelled along the corridor. From these trajectories, vehicle-pedestrian conflicts were recognised and measured using the conflict indicator known as post-encroachment time, which refers to the time difference between the first road user exiting an encroachment zone (the area where two road users have a possibility of a conflict) and a subsequent road user entering the same encroachment zone. Several covariates characterising vehicle-pedestrian interactions were also extracted from the data to introduce non-stationarity to the Extreme Value Theory models. Both Block Maxima and Peak Over Threshold sampling-based models were estimated in the Bayesian framework. The estimated models were used to interpret the interplay of the crash risk and various covariates, such as pedestrian speed, vehicle speed, pedestrian count, and vehicle count, which can help to develop effective strategies to reduce crashes and enhance transportation safety on a larger scale.

Models using both Block Maxima and Peak Over Threshold extreme sampling techniques were found to estimate the observed crashes well. Notably, the Block Maxima sampling-based model was found to be more accurate than Peak Over Threshold sampling-based model based on mean crash estimates as well as confidence intervals. The findings of this research highlight the need for a holistic approach to road safety analysis, considering the entire corridor rather than isolated segments. Autonomous vehicle sensor data can be used to identify pedestrian crash zones on a transport network.

Table of Contents

List of Figures

List of Tables

List of Publication

1) Singh, S., Ali, Y. and Haque, M.M. (2023). Assessing corridor-wide pedestrian safety using autonomous vehicle sensor data: A Bayesian extreme value theory modelling framework, *Accident Analysis and Prevention, Under review*

List of Abbreviations

Abbreviation	Full Form
AV	Autonomous Vehicle
CRSS	Crash Report Sampling System
DIC	Deviance Information Criterion
DOC	Deceleration Occurrences caused by Conflicts
DRAC	Deceleration Rate to Avoid Collision
FCA	Fuzzy Cellular Automaton
GES	General Estimates System
MDV	Manually Driven Vehicle
MTTC	Minimum Time to Collision
OLS	Ordinary Least-Squares
PET	Post-Encroachment Time
PVCA	Pedestrian Vehicle Conflicts Analysis
RC	Risk of Crash
RLC	Return Level of Cycle
SEM	Structural Equation Modelling
SSAM	Surrogate Safety Assessment Model
TDTC	Time Difference To Collision
TTC	Time To Collision
VKT	Vehicle Kilometres Travelled
VMT	Vehicle Miles Travelled
VRU	Vulnerable Road User
WMT	Walking Miles Travelled

Statement of Original Authorship

This thesis represents original work that has not been previously submitted for any academic recognition at this or any other educational institution. To the best of my knowledge and belief, I affirm that the thesis does not include any previously published or authored material by any other individual unless proper credit and acknowledgment have been given through appropriate references.

Acknowledgement

I would like to express my sincere gratitude and appreciation to a number of individuals who have contributed to my Master of Philosophy degree. Their guidance, assistance, and encouragement have been invaluable throughout this journey.

First and foremost, I extend my deepest thanks to my principal supervisor, Dr Shimul Haque, for his unwavering commitment and support with his domain expertise. His guidance, constructive feedback, and scholarly insights have shaped the direction and quality of my research. His patience, consistent supervision, accessibility, and willingness to share his knowledge have been instrumental in my academic growth and development.

I am also grateful to my associate supervisor, Dr Yasir Ali, for his valuable contributions and insightful suggestions. His expertise and critical evaluations have immensely enriched the depth and breadth of the research outcome. I am truly grateful for his time, dedication, and the opportunity to learn from his vast knowledge.

Furthermore, I would like to acknowledge the support and guidance provided by the faculty members and staff of the Queensland University of Technology. Their commitment to academic excellence and willingness to assist students in their academic pursuits have helped create a conducive learning environment.

I am deeply indebted to my fellow researchers and colleagues, whose valuable discussions and feedback have played a vital role in shaping my ideas and refining my research methodology. Their willingness to share their expertise and engage in scholarly discussions has been inspiring and enriching.

My heartfelt appreciation goes to my family, my wife Kirti, parents Dr Ishwar Singh and Dr Sunita, and my sister Dr Priya for their unwavering support, love, and understanding throughout this demanding academic endeavour. Their encouragement, belief in my abilities, and sacrifices have been the driving force behind my accomplishments. I am grateful for their patience during long hours of study, understanding during periods of stress, and continuous motivation for me to pursue excellence. Their presence, kind words, and belief in my capabilities have provided much-needed strength and motivation during challenging times.

I would like to communicate my sincere appreciation and thankfulness to the Australian Government for providing the necessary scholarship to make this body of research possible. An Australian Government Research Training Program Scholarship supported the work and the outcomes of this research.

Last but not least, I would like to express my gratitude to the broader academic community and the numerous authors whose research and publications have contributed to the foundation of knowledge in my field of study. Their pioneering work has been an invaluable resource throughout my research journey.

I am humbled and deeply grateful for the invaluable contributions of all the individuals mentioned above. Their support, guidance, and encouragement have played an instrumental role in my Master of Philosophy degree. Thank you all for your unwavering support and belief.

This page intentionally left blank.

Chapter 1 Introduction

1.1.Background

Road safety analysis is a critical aspect of the transportation engineering discipline that examines road-user interactions and works towards reducing the frequency and severity of road crashes. The importance of road safety analysis becomes quickly apparent when looking at the global cause of death statistics. Road crashes are the global leading cause of death in children aged between 5 and 14 and one of the top three causes for the age group of 15 to 49 (IHME, 2021). With the exceedingly high number of lives lost in road crashes, increasing number of vehicles on the roads every year, and the growing complexity of transportation systems, addressing road safety concerns has become a pressing issue worldwide.

Traditionally, road safety analysis has been focused on reactive methods, which involve studying individual crash sites or specific intersections after crashes occur. These studies rely on crash data and historical records to identify underlying patterns, contributing factors and devise remedial measures. While such an approach has provided valuable insights into the causes of crashes and guided local safety interventions at specific locations, they often remain limited in their effectiveness in mitigating fundamental causes of crashes due to shortcomings of historical crash datasets. These historical crash datasets often require large-scale data collection programmes to gather sufficient information for meaningful analysis, which can take years' worth of time to accumulate (Wu and Xu, 2018). Also, researchers have often reported that crash data suffer from under-reporting, limited sample size, and unobserved heterogeneity (Ali et al., 2023a).

The reactive approach to road safety analysis poses several other challenges and limitations. Firstly by design, it emphasizes the remediation of a problem after incidents have already taken place rather than taking preventive measures to avoid crashes altogether. This reactive nature of analysis leads to a cycle of blackspot location identification, countermeasure implementation, and a waiting period for the next blackspot to appear before further actions can be taken. Therefore, this approach may overlook the underlying causes of crashes and fail to break the above-mentioned cycle. Secondly, a crash data-based reactive road safety approach often fails to capture the complex interactions and interdependencies of traffic movements by multiple road users along transportation corridors. A transport network is an interconnected system of roads with diverse

characteristics, including varying traffic volumes, trip purposes, roadway geometries, environmental conditions, and driver behaviour. Road safety analysis on an isolated segment or at a specific intersection neglects the broader influence of corridor-related contributing factors. This limitation restricts the effectiveness of the implemented countermeasures, as they may not address the underlying systemic issues leading to crashes along the corridor. Moreover, the reactive nature of road safety analysis leads to inefficient allocation of resources. With limited resources and a lack of a comprehensive understanding of the crash risk factors across an entire corridor, transportation agencies face challenges in identifying and prioritizing high-risk locations solely based on historical crash data, often resulting in suboptimal resource allocation and potential missed opportunities for proactive safety interventions.

In recent years, the limitations of the reactive road safety approach have spurred the development of proactive safety analysis methods. Proactive approaches aim to identify potential road safety issues before they lead to crashes, allowing for much quicker interventions and prevention strategies. Examples of proactive safety analysis approaches beyond others include predictive models, simulation techniques, and proactive safety audits to assess the safety performance of transportation corridors and identify crash risk areas. While proactive safety analysis techniques show promise in their ability to anticipate and mitigate crash risks, they are not without their own limitations.

Predictive models utilise historical crash data and statistical techniques to forecast future crash frequency and identify blackspot areas. However, these models heavily rely on the assumptions determined from analysing past crash trends. These assumptions may not hold true if there are significant changes in traffic demand, infrastructure, or driver behaviour.

Simulation models, on the other hand, simulate traffic operations and interactions to assess road safety performance. These models allow for scenario testing and evaluation of different network design options or safety interventions. However, simulation models require massive data inputs and extensive calibration to produce reliable results. The quality and availability of data, such as traffic volume and driver behaviour, can significantly impact the accuracy and applicability of the model results.

Moreover, both predictive and simulation models have limitations in considering the human element of road safety. Driver behaviour, perception, and decision-making are

inherently complex and challenging to model accurately. The variability in human behaviour and the influence of factors such as distraction, fatigue, and impairment pose significant challenges in predicting and simulating road safety outcomes.

One prominent proactive safety framework in particular, based on traffic conflict information, has gained much popularity in the past decade. This approach, which uses traffic conflicts as spatial/temporal proximity of crashes, was first conceptualised by Swedish researchers in the late 1970s. Hydén (1987) expanded on that work and categorised different scenarios observed in a typical traffic stream into Hyden's Safety Pyramid conceptual framework. Hyden's Safety Pyramid establishes a link between traffic conflicts and traffic crashes and provides a theoretical foundation for valuable proactive safety analysis tools. It illustrates the hierarchical relationship between various types of road events, ranging from undisturbed traffic operations, near-miss incidents, and conflicts to actual crashes with varying degrees of severity, as illustrated below in Figure 1.1 According to Hyden's Safety Pyramid, near-miss incidents and conflicts are precursors to crashes and occur more frequently than crashes, making them reliable indicators of potential safety issues. Conflict-based studies can capture near-miss incidents, which serve as valuable indicators of potential crash hotspots and can inform measures to improve safety. Emerging technologies such as autonomous vehicles might introduce new types of incidences and challenges. These crashes could still be influenced by human factors and the mass scale adoption of the technology is still long way away. Additionally, the fundamental relationship between crashes, conflicts and road operations will remain unchanged. Hence, Hyden's theoretical framework is well suited to assess emerging vehicle technologies such as autonomous vehicles (Ali et al., 2023a).

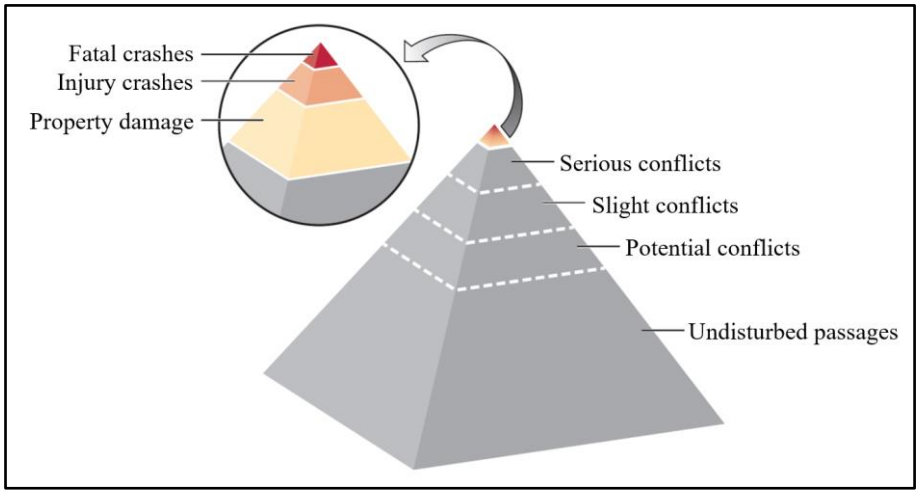


Figure. 1.1. Hyden's safety pyramid (Hydén, 1987).

Over time, several statistical models have been developed to mathematically explore the correlation between conflicts and road crashes, as illustrated by Hyden's safety pyramid. These include causal models, probabilistic models, and extreme value models, among others. For this study, the focus will be on utilizing extreme value models due to their unique advantages. Extreme value models are particularly suitable for analysing rare events with severe consequences, such as financial risk, natural disaster risk, and crash risk. From the context of the study, the models allow the identification of high-risk locations or situations that may not be apparent from traditional reactive approaches. To do so, extreme value models provide insights into the tail distribution of conflicts, capturing the infrequent road conflict events to predict potentially catastrophic crashes and inform safety interventions. Notably, conflict processing for extreme value theory analysis has made significant advancements in recent years. Researchers have developed robust video analytics tools for conflict extraction to establish a relationship between conflicts and crashes (Ali et al., 2023b; Arun et al., 2022; Arun et al., 2021a; Zheng et al., 2018). While conflict-based analysis provides significant insights into crash risk frequency and severity, collecting and analysing conflict data can be labour-intensive and time-consuming, making it less feasible for large-scale implementation.

The integration of emerging technologies, such as autonomous vehicle sensor data, can significantly contribute to conflict-based safety analysis. Different from labour-intensive conventional conflict data collection techniques requiring constant oversight, autonomous vehicles are equipped with a wide array of sensors, including LiDAR, radar, and cameras. These sensors continuously monitor the surrounding environment and capture detailed information about road users, objects, and potential conflicts. By leveraging this rich sensor data, transportation professionals can gain a more comprehensive understanding of interactions and conflicts between autonomous vehicles, pedestrians, cyclists, and conventional vehicles. The detail-oriented and real-time nature of autonomous vehicle sensor data generates scope for an unprecedented level of precision in conflict-based safety model outcomes. It enables the detection of subtle conflicts and near-miss events that may go unnoticed by human observers or traditional data collection methods. Furthermore, as autonomous vehicle technology advances and more autonomous vehicles share the road, the collective data from the sensors of these vehicles can provide valuable insights into overall traffic patterns, emerging conflict hotspots, and potential areas for road safety improvement. This study contributes to the ongoing progress in conflict-based extreme

value theory models by building upon these developments and advancing the methodology in the large-scale corridor-wide application of pedestrian safety analysis.

1.2.Research motivation and needs

After identifying the limitations of current crash risk analysis techniques, further research was conducted on global road safety statistics to focus on a specific research question as a part of this study. Looking at the crash data collected by multiple international organizations and collated by World Health Organization, more than half of all road network-related deaths are among vulnerable road users: pedestrians, cyclists, and motorcyclists (WHO, 2022). Of all the vulnerable road users, pedestrians face a much higher crash risk in the current car-dominant road environment (Khayesi, 2020). This crash risk is hypothesised to be differential with advancements in vehicle technologies such as connected vehicles (Ali et al., 2022b). Even though pedestrians represent a fraction of kilometres travelled on the road network, they still account for more than one in five road fatalities globally (WHO, 2023). In general, the average per kilometre pedestrian fatality risk is nine times higher compared to a car occupant (Job, 2020). With that differential risk in mind, this research focuses on understanding the pedestrian crash risk by applying corridor-wide conflict-based safety analysis techniques using autonomous vehicle sensor data.

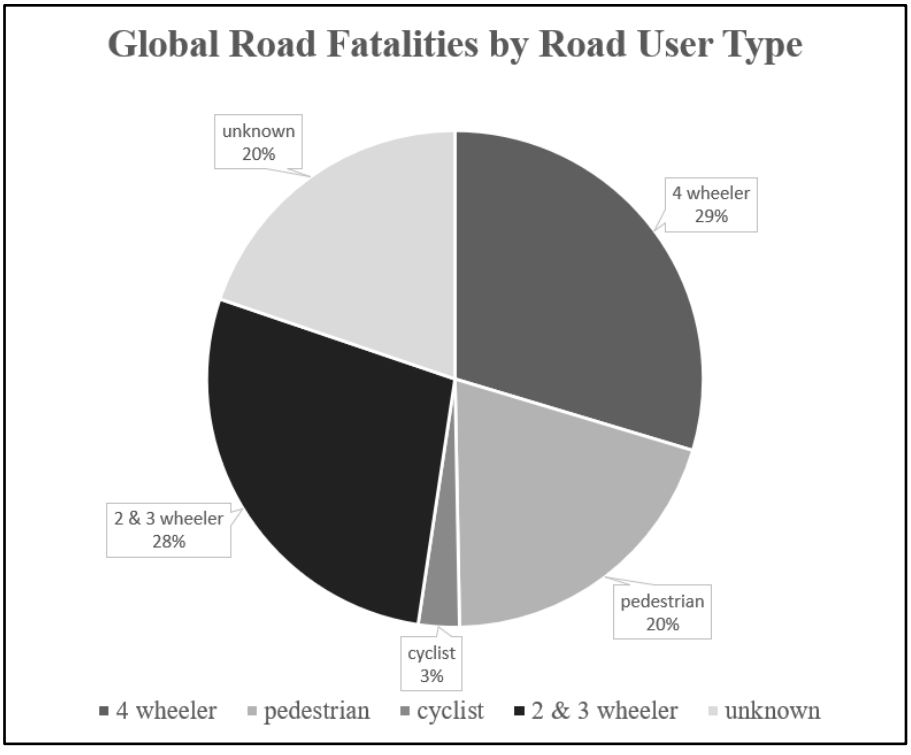


Figure. 1.2. Global road fatalities by road user type in 2016 (WHO, 2023).

Pedestrian trips encompass multiple road sections, such as intersections and mid-blocks, that collectively form a corridor. They have very high network access and mobility and often exhibit unpredictable, erratic, and haphazard movements. This behaviour challenges the effectiveness of the traditional spot-safety approach, which primarily focuses on intersections, as it may fail to identify the true underlying pedestrian crash risk on the corridor. To that extent, an analysis of crash records in the state of Queensland, Australia, reveals that a significant proportion of pedestrian fatalities over the past two decades, around 78%, occurred away from intersections (TMR, 2022). Recognizing a clear need for a more comprehensive pedestrian crash risk assessment strategy, our research question aims to investigate how corridor-wide conflict-based safety analysis, independent of historical crash history, can better assist with it.

Despite their potential benefits, conflict-based safety studies have seen limited scale application in practice. One major reason is the resource-intensive nature of data collection and processing required for such studies. For instance, conducting field observations or recording videos to capture conflicts often involves significant time, personnel, and financial investments. Installation of video cameras at multiple locations, data collection over a period of time, and then data processing requires significant equipment, workforce, and computing power. This resource-intensive process can be a barrier to widespread implementation, especially for large-scale studies covering extensive road networks. In this regard, autonomous vehicle data holds promise as a potential substitute for traditional data collection techniques. Autonomous vehicles are fitted with an array of sensors, such as LiDAR, radar, and cameras, which continuously capture detailed information about their surrounding environment, including pedestrians, cyclists, and other vehicles. This rich sensor data can be leveraged to track objects, identify and analyse conflicts, providing a cost-effective and efficient alternative to manual data collection. By utilizing autonomous vehicle data, researchers can potentially overcome the resource limitations associated with traditional methods and scale up conflict-based safety studies to a corridor-wide level. Very recently, several empirical autonomous vehicle datasets such as KITTI (Geiger et al., 2013), Argo (Wilson et al., 2021), Lyft (Kesten et al., 2019), Waymo (Sun et al., 2020), nuScenes (Caesar et al., 2019) have been made publicly available providing access to autonomous vehicle sensor data that can help us to understand traffic safety better. Interestingly, despite the potential advantages of using autonomous vehicle sensor data for conflict-based safety analysis, there is a notable absence of studies exploring these datasets'

application at a corridor-wide scale. Thus, there is a clear research gap and opportunity to conduct a corridor-wide conflict-based pedestrian safety analysis study utilizing autonomous vehicle sensor data.

The following inference can be drawn from the above discussion regarding pedestrian safety analysis. First, crash-based statistical models have been frequently used in literature for analysing pedestrian safety either at mid-blocks or signalised intersections. These models predominantly utilise police and government authority-reported data, with well-known limitations such as crash under-reporting and limited availability of behavioural information. Second, recognising issues with police-reported data, extreme value theory models have been developed to assess pedestrian safety using traffic conflicts. Third, several data collection methods have been used to obtain traffic conflicts, with video data predominantly used at one or multiple scattered intersections. Finally, recent vehicle technologies such as autonomous vehicles present unprecedented opportunities to better understand traffic safety, and a few applications related to pedestrian safety can be found. However, our understanding remains elusive about fully leveraging the capabilities of autonomous vehicle sensor data for analysing pedestrian safety at a corridor level, and this data has not received due attention in the literature. This research gap motivates the present study.

1.3.Research objectives and questions

This research proposes a framework to estimate extreme value theory-based corridor-wide pedestrian crash risk model using autonomous vehicle sensor data. In order to fulfil the aim, Block Maxima and Peak Over Threshold extreme sampling techniques are used.

Based on the research gaps identified through the literature review and the aim of the study outlined above, this study aims to answer the following research questions:

- RQ1: How to process and interpret autonomous vehicle sensor data in a meaningful way?
- RQ2: How to extract road user conflict information from autonomous vehicle sensor data and check the quality of the output?
- RQ3: How can the Extreme Value Theory modelling technique be applied to a corridor-wide safety assessment framework?

To effectively address the research questions and achieve the overarching goal of this study, the following tasks need to be accomplished:

- Task 1: To develop autonomous vehicle sensor data processing methodology for extraction of road user trajectories and conflict indicators (corresponds to RQ1 and RQ2)
- Task 2: To propose a corridor-wide crash risk analysis framework based on Extreme Value Theory (corresponds to RQ2 and RQ3)

Research contribution

The study contributes to the literature by establishing the efficacy of extreme value models for corridor-wide safety studies. By utilizing these models, the study intends to demonstrate their effectiveness in capturing rare and severe events, thereby enhancing our understanding of high-risk locations along a corridor. Second, by utilising extensive autonomous vehicle sensor data, the study offers a unique opportunity to explore conflict-based safety analysis on a larger scale, which has been identified as a research gap in a review study by Ali et al. (2023a). Also, the framework offers the following benefits: 1) it facilitates the identification and prioritization of crash-prone locations, enabling targeted interventions to prevent crashes before they occur, and 2) it utilises extensive autonomous vehicle sensor data to enhance our understanding of conflicts and safety dynamics and efficient detection of crash-prone locations on a transport corridor.

1.4.Thesis outline

This thesis is structured to comprehensively explore the research topic and present a logical progression of ideas.

Chapter 2 serves as the literature review, offering a comprehensive survey of existing studies and the progression from crash-based to conflict-based safety analysis. This chapter aims to establish a strong theoretical foundation and provide a clear context for the research gap in the existing literature and the research contribution of this study.

Chapter 3 draws a brief outline of the methodology and framework used in the study. Following that, the chapter introduces extreme value modelling techniques in the context of this study. It elaborates on the mathematical underpinnings of extreme value modelling and demonstrates its application in analysing conflict data derived from autonomous vehicle datasets. It showcases the methodology employed in the study, including the statistical models used and any specific considerations or adaptations made for the corridor-wide analysis. Overall, the chapter lays the theoretical foundation for the following chapters and how they link back to the study methodology.

Chapter 4 focuses on the autonomous vehicle datasets considered in the study. It delves into the selected dataset's details, its characteristics, data collection methods, and any pre-processing steps taken to ensure data quality. This chapter serves as a crucial link between the theoretical framework established in Chapter 3 and the empirical analysis conducted in the following chapters.

In Chapter 5, the focus shifts to the evaluation of the calibrated model's performance. This chapter presents the results and findings obtained from the application of extreme value models to conflict data. It assesses the model's ability to capture and predict rare and severe events, providing insights into high-risk locations along the corridor.

Finally, Chapter 6 concludes the thesis by summarizing the key findings, discussing their implications, and highlighting the contributions of the study to the field of road safety analysis. This chapter also presents recommendations for future research and suggests potential avenues for the practical application of the research outcome.

Chapter 2 Literature review

Given the focus of the research question for this study is pedestrian safety, this section provides a concise overview of the pedestrian safety literature, with more comprehensive reviews available in earlier studies. For detailed information on specific topics, readers are directed to separate reviews such as those focused on traffic conflict measures (Arun et al. (2021b) and Arun et al. (2021c)), general traffic conflict modelling (Zheng et al., 2021), and extreme value theory models (Ali et al., 2023a).

2.1. Crash-based pedestrian safety studies

Crash-based pedestrian safety studies analyse traffic crash data to identify factors contributing to pedestrian injuries and fatalities. Pedestrian safety studies typically gather crash data through extensive data collection programs or use existing crash data to develop statistical regression models to evaluate pedestrian crash safety. In past studies, multiple statistical regression models such as the chi-square statistical test, least square regression analysis, and binary logistic regression have been used to develop crash-based pedestrian safety models. The following section highlights a representative sample of studies that use crash-based safety analysis followed by a few critical limitations of crash based techniques which led to development of conflict based safety analysis techniques.

Ammar et al. (2022) developed a chi-square statistical test-based model to study and examine pedestrian crash data from the United States. The study focused on serious pedestrian injury and fatality instances at intersections for a decade. The crash data was collected at intersections from the General Estimates System (GES) for 2013-2015 and the Crash Report Sampling System (CRSS) for 2016-2018. The study discovered that the overall pedestrian crash risk increased from 2013 to 2018. Moreover, the study analysed identified key factors influencing pedestrian crash severity. Two logistic regression models were used, and 14 explanatory variables were considered. The results revealed that pedestrian age, lighting condition, vehicle body type, and vehicle pre-crash movement were significant factors in both datasets. Older pedestrians had a higher likelihood of severe or fatal injuries compared to children. Crashes occurring during dark lighting conditions had a significantly higher probability of serious or fatal outcomes. Light trucks and buses were associated with a higher risk of severe or fatal crashes than passenger cars. Additionally, the study found differences in the impact of some factors between the two datasets, such as the age group of 56-65, vehicle pre-crash manoeuvres, and pedestrian actions prior to the crash. The findings of the study

suggest the need for interventions to improve pedestrian safety at intersections, including infrastructure improvements, regulations, and increasing visibility for both pedestrians and drivers, especially during nighttime conditions.

In another study in the state of Texas, US, Bernhardt and Kockelman (2021) analysed the factors leading to pedestrian crashes using an ordinary least-square (OLS) regression analysis to formulate right-of-way measures to mediate those contributing factors. The study analysed pedestrian crash counts, and pedestrian deaths per vehicle miles travelled (VMT) and per walking miles travelled (WMT) in Texas counties from 2010 to 2018. The models incorporated demographic, climate, and roadway factors across 254 counties using data from databases such as CRIS database, United States Census Bureau, Texas Association of Counties, and PRISM Climate Group. The use of ordinary least-squares regression allowed for efficient processing and comprehension of large amounts of data at the county level. VMT and WMT measures were employed to normalize crash counts and control for the scale effects. The study examined the influence of speed, darkness, distracted drivers and pedestrians, the presence of signals and crosswalks, climate, and homelessness on pedestrian crashes. The findings revealed that higher speeds and darkness increased the severity of crashes, while the presence of pedestrian facilities and appropriate crossing behaviour reduced crash rates. Climate and weather, such as warmer temperatures and precipitation, were associated with increased pedestrian activity and crash rates. Additionally, homelessness emerged as an important factor in pedestrian crashes, warranting further attention in pedestrian safety discussions.

Gooch et al. (2022) aimed to identify risk factors associated with severe pedestrian crashes on specific road segments in Massachusetts through a systematic safety analysis. The researchers used geolocated crash data from the Massachusetts Registry of Motor Vehicles and roadway data from the Massachusetts Road Inventory to focus on midblock locations and included driveways, non-junction areas, and unknown or unreported junction types. Binary logistic regression models were developed for each facility type to assess the probability of severe pedestrian crashes based on various independent variables. The results showed several significant risk factors across different facility types. Variables such as the number of lanes, traffic volume, and the presence of a median/barrier were positively correlated with the probability of severe pedestrian crashes. Higher employment density, population density, and a higher proportion of employment in certain industries were also associated with an increased likelihood of crashes. Additionally, lower median household income and the presence of environmental justice indicators (indicating socioeconomic disparities) were positively

correlated with crash probability, highlighting equity issues in pedestrian safety. Transit stop presence and density were positively associated with the likelihood of severe pedestrian crashes.

Escobar et al. (2021) estimated that over 20% of the road-crossing manoeuvres could be classified as potential traffic conflicts in a pedestrian road-crossing behaviour analysis study using logistical regression modelling. The study focused on analysing pedestrian behaviour at critical points in Manizales, Colombia, to understand the factors contributing to road incidents. The researchers conducted direct observations at seven selected points, including road intersections and mobility corridors, and recorded various aspects of pedestrian behaviour, such as crossing behaviour, distractions while crossing, interaction with traffic, and crossing speed. A total of 33,561 pedestrians were observed during the study, and it was found that 65.19% of pedestrians crossed at designated locations, while the rest engaged in risky behaviours such as crossing at non-designated locations or diagonally. Children exhibited the highest proportion of risky behaviour, while older adults tended to use diagonal trajectories. The analysis also revealed variations in pedestrian behaviour based on the specific locations analysed, network typology, and intersection type. Furthermore, pedestrians were frequently observed being distracted while crossing, with talking to other pedestrians and using headphones being the most common distractions. The study also examined pedestrians' interaction with motorized road actors, with crossing in low traffic conditions being the most prevalent behaviour. The findings highlight the importance of considering pedestrian behaviour, age groups, and specific location characteristics when designing road safety interventions.

Peng et al. (2020) analysed pedestrian crashes and the corresponding contributing factors using a structural equation model. This study used structural equation modelling (SEM) to analyse pedestrian paths and investigate various factors directly or indirectly affecting the injury severity in vehicle-pedestrian crashes at mid-blocks. The study utilized two models: a multinomial logit model to examine the effects of variables on the pre-crash behaviour of pedestrians and an ordered logit model to uncover the associations between injury severity and contributing factors, including pre-crash behaviour. The analysis revealed several significant findings. The age of pedestrians was found to be correlated with pre-crash behaviours, with older pedestrians less likely to engage in rushing or running into the road but more likely to conduct improper crossings. The number of lanes and environmental conditions, such as nighttime and wet surface conditions, also influenced pre-crash behaviours. Regarding injury severity, certain pre-crash behaviours increased the likelihood of injuries with varying severity

ranging from non-incapacitating, non-incapacitating evident injuries to incapacitating and fatal injuries. Vehicle type, the first point of impact, and speed limit were significant factors affecting injury severity, with heavier vehicles and impacts on the sides of the vehicle increasing the likelihood of more severe injuries. As mentioned above, the study also calculated the indirect effects of explanatory variables on injury severity through pre-crash behaviours. The results showed that older pedestrians had a higher likelihood of severe injuries, and higher speed limits and nighttime conditions increased the probability of more severe injuries. The findings suggest the importance of considering both direct and indirect effects of factors on injury severity to improve pedestrian safety in mid-block crashes.

Different from the statistical studies above, there are several descriptive studies to test the influence of certain explanatory variables on pedestrian crash risk. To this end, Bendak et al. (2021) aimed their study at understanding the impact of various contributing factors on pedestrians' road crossing behaviour using Chi-squared modelling. The study aimed to analyze pedestrian behaviours at signalized crosswalks in Sharjah, UAE, using direct roadside observations. Data on various pedestrian behaviours, socio-demographic factors, and road details were collected at ten crosswalks. A total of 708 pedestrians were observed, and their behaviours were recorded. The average walking speed was 1.22 m/s, slightly faster than reported in previous studies. Gender, age, day of the week, the number of people waiting, type of pedestrian lights, walking with children, temperature, green light duration, and the number of lanes were found to have significant effects on pedestrian behaviours. Males were more likely to chat and cross on red, while females walked slower. Pedestrians between 16 and 39 years of age were more likely to use mobile phones when crossing and also cross on the red signal. Walking speed and looking around before crossing were affected by the day of the week. Pedestrians crossing with larger groups walked slower. Mid-block crosswalks had higher rates of looking around and faster walking speeds than road intersection crosswalks. Pedestrians walking with children had slower walking speeds and were less likely to use mobile phones. Pedestrians walked faster, crossed on red more often, and used mobile phones less frequently at higher temperatures. Crosswalks with shorter green times and longer red times had higher rates of walking outside designated areas and crossing on red. Pedestrians were more likely to cross on red at crosswalks with fewer lanes. Pedestrians crossing on red walked faster, and those carrying loads walked slower than others. In the study, over seventeen per cent of the pedestrians at least partially undertook the road-crossing manoeuvres during the red signal, creating a higher probability of a conflict. The findings provide valuable insights into

pedestrian behaviour and can inform the development of strategies to improve pedestrian safety.

The studies mentioned above primarily focused on individual or handful signalised intersections or mid-blocks, with limited attention given to pedestrian safety analysis at a corridor level. However, there is a need to examine pedestrian safety within the context of entire corridors. (Hong et al., 2016) introduced a novel approach by developing a spatially autoregressive and heteroskedastic space-time pedestrian exposure model. The study employed a methodology that included tests for spatial dependency, endogeneity, heteroscedasticity, density, and time occupancy. This model incorporated spatial lags and endogenous network topologies, capturing the stochastic network design effects in estimating pedestrian safety. Their work aimed to provide a more comprehensive understanding of pedestrian safety by considering the complexities of the entire corridor rather than individual isolated points of analysis. The results of the tests confirmed the presence of spatial autocorrelation, indicating that crosswalks in downtown Seattle exhibit similar characteristics in terms of density and time occupancy. The study addressed the issues of endogeneity and weak instruments by using instrumental variables, conducting various tests, and accounting for heteroscedasticity in developing the spatial model. The proposed SARAR model with exogenous and endogenous regressors accounted for these statistical issues. The results of the SARAR models showed that endogenous network topology measures and other exogenous variables significantly influenced pedestrian area density. Factors such as network connectivity, proximity to other crosswalks, and trip generation volumes from nearby facilities were found to be significant. The models also revealed positive spatial effects, indicating the influence of neighbouring crosswalks on density. The study provided insights into the relationship between network characteristics and pedestrian density, highlighting the importance of considering spatial dynamics in urban planning and design.

Despite the simplicity of crash-based regression analysis and its strength in establishing causal relationships and providing detailed insights into pedestrian safety, there are several limitations associated with this approach. Issues such as limited accessibility to crash data (Ismail et al., 2011) and data quality concerns, including geographical imprecision, underreporting, and data logging errors (Zheng et al., 2021), hinder the reliability and accuracy of the models. Furthermore, due to the relatively low frequency of pedestrian-vehicle crashes compared to other types of crashes, there has been a growing recognition of the need for alternative approaches to analyse and improve pedestrian safety. As a result, there is a growing

interest towards conflict-based safety analysis methods (Arun et al., 2021b), which offer a different perspective by focusing on traffic conflicts as indicators of potential crash risk rather than solely relying on crash data.

2.2.Conflict-based pedestrian safety studies

Looking at the literature available in the field of proactive conflict-based safety, several techniques have been devised to assess vehicle-to-vehicle conflicts in past studies. For example, Zheng and Sayed (2020) applied the Extreme Value Theory framework and proposed a real-time crash risk prediction approach for signalized intersections.

On the other hand, when it comes to modelling vehicle-to-pedestrian conflicts, the literature offers relatively fewer modelling techniques. Among the modelling techniques employed for vehicle-to-pedestrian conflicts, some notable approaches are listed below:

Guo et al. (2020) developed an extreme value theory model with a peak-over-threshold extreme sampling technique designed for before-and-after safety assessments. The study quantified the effects of implementing Leading Pedestrian Intervals (LPIs) at two intersections in the central business district in Vancouver to enhance pedestrian safety. Leading pedestrian interval is additional green time given to pedestrians to cross an intersection before left-turning vehicles start moving, reducing conflicts between pedestrians and vehicles. The study collected data before and after the implementation of LPI in early July 2018. Control sites with similar characteristics and geographic proximity were chosen to account for unobserved bias. Video data collection periods were kept consistent for both before and after intervention periods to ensure consistency. The data were analyzed using automated traffic conflict analysis techniques. The study used the Peak Over Threshold (POT) approach based on the Generalized Pareto Distribution to estimate treatment effects. A hierarchical Bayesian model was used to estimate Generalized Pareto Distribution parameters, and the Deviance Information Criterion was used to select the best-fitted model. The results showed a significant reduction in extreme-serious conflicts, ranging from 18.1% to 20.9%. This reduction indicates improved safety after implementing mitigation measures. The study highlighted certain limitations, including the small size of the dataset and the need for additional research on varying environments and long-term effects.

Sun et al. (2022) applied the game theory framework. The study conducted a field survey to analyze pedestrians' traffic characteristics and crossing behaviours on unsignalised crosswalks. For the study, three typical unsignalised sections with crosswalks were selected in

Beijing, China. The study found that the average walking speed of pedestrians crossing the street was 1.25 m/s, with the peak concentration at 1.1-1.3 m/s. In this study, pedestrian age had the greatest influence on walking speed, with slower speeds observed in older pedestrians. Male pedestrians had higher walking speeds than females, and faster vehicle speeds led to faster pedestrian crossing speeds. The study also found that pedestrians often needed to pause multiple times to safely cross the street, with about 15% of pedestrians pausing three or more times. The waiting time for crossing was influenced by the number of pauses, with 70.18% of waiting times falling in the range of 10-30 seconds. Pedestrians became less risk-averse as the waiting time increased, leading to behaviours such as rushing to cross the street. Risk assessment and waiting delay were important factors influencing pedestrian crossing decisions, with pedestrians judging the distance from vehicles, crossing pace, acceptable crossing gap, and crossing mode to make their decisions. Pedestrian and driver behaviour characteristics and the psychological characteristics of pedestrians when crossing the street were also analysed. The findings of this study provide valuable insights for understanding and addressing the challenges associated with unsignalised pedestrian crossings.

Ghadirzadeh et al. (2022) used the multinomial logit framework. The study examined pedestrians' crossing behaviours and risk-taking tendencies using two conflict indicators: Post-encroachment Time (PET) and Time to Collision (TTC). PET represents the time difference between the first road user leaving the conflict point and the second road user reaching that point, while TTC indicates the expected time for two road users to collide if their current speed and direction remain unchanged. The research selected five pedestrian crossings in Qazvin, Iran, and collected data through video recordings. A total of 752 pedestrians were analysed, and their characteristics and behaviours were extracted. Based on the TTC and PET indexes, the study used a binary logit model to estimate pedestrians' risk-taking behaviours. The results showed that running pedestrians, the presence of companions, and weather conditions significantly influenced risk-taking tendencies. Additionally, pedestrians' age and gender also showed differences in crossing behaviours. The study concluded that pedestrians' risk-taking behaviours are influenced by various factors, including their individual characteristics, environmental conditions, and interactions with other road users. The findings contribute to a better understanding of pedestrian behaviours and can inform the design of safer pedestrian crossings.

Li et al. (2021) used the fuzzy cellular automata framework. The study conducted a simulation experiment to analyse the safety and efficiency of non-signalised midblock

crosswalks. Two key factors, proportions of obedient drivers and traffic flow rate, were investigated. Safety was measured using the Time Difference to Collision (TDTC) indicator, which represents potential conflicts between pedestrians and vehicles. Efficiency was evaluated using the Deceleration Occurrences caused by Conflicts (DOC) indicator, which indicates the frequency of conflicts between pedestrians and vehicles. The simulation results showed that stricter traffic rules led to fewer serious conflicts but increased vehicle-stopping occurrences, reducing traffic efficiency. Higher traffic flow rates increased conflict severity, while pedestrian flow rates had minimal impact on efficiency. Based on the findings, two policy insights were provided: promoting drivers' yielding behaviour and suggesting the installation of midblock zebra crosswalks in areas with frequent unauthorized midblock crossings. The study identified a minimum pedestrian flow rate of 290 per hour for installing midblock zebra crosswalks. The research concluded that the Fuzzy Cellular Automaton (FCA) model effectively represented pedestrian-vehicle interactions and provided valuable insights for traffic management at non-signalized midblock crosswalks. Future research could consider incorporating pedestrians' swerving behaviours, vehicles' lane-changing behaviours, and different yielding behaviours of vehicle types to enhance the realism of the traffic conditions.

Santhosh et al. (2020) used the Pedestrian Vehicle Conflict Analysis Framework. Their study aimed to examine conflicts between vehicles and pedestrians using a video-graphic method and both manual and simulation-based data analysis techniques. The study areas were chosen based on high vehicular and pedestrian volumes during peak hours, making the findings applicable to similar locations in Asian countries. Data collection involved capturing factors related to pedestrians, vehicles, and collisions through field studies and video-graphic surveys. The analysis focused on volume and speed data to determine the causes of conflicts, employing the Pedestrian Vehicle Conflicts Analysis (PVCA) method, which involved three steps: conflict identification, classification based on severity factors, and determination of conflict severity grade. The study found that two-wheelers and male adults constituted the majority of the volume in both intersections studied in the research. The conflicts were classified as slight and serious, with more serious conflicts occurring during morning peak hours due to the aggressive nature of pedestrians. Pedestrian-vehicle simulation using VISSIM software was conducted to analyse conflicts and recommend countermeasures. The simulation results showed a good fit with observed data, and the Surrogate Safety Assessment Model (SSAM) was used to validate the conflicts obtained from the field. To reduce conflicts, rerouting pedestrian volumes and restricting crossing paths were suggested as mitigation measures. A comparison between the

unsignalised T and X intersections revealed that converting an uncontrolled intersection into a controlled intersection reduced conflicts. The study concluded that conflicts at unsignalised intersections depend on pedestrian and vehicular volumes, and a combination of manual and simulation methods provides detailed insights into conflict causes and severity. The future scope of the work includes assessing conflict severity solely through simulation-based techniques.

The above models briefly highlight some of the conflict-based pedestrian safety models that have been used in the past literature. However, for the purposes of this study, the focus lies specifically on utilizing extreme value theory. The following section will delve deeper into the past literature focused on the application of the Extreme Value Theory in conflict-based pedestrian safety studies. The following section will highlight contributions from some representative studies to shed light on the existing research regarding pedestrian safety analysis using extreme value theory.

Fu and Sayed (2021) focused on developing a crash estimation method using traffic conflict indicators. Three conflict indicators, namely Minimum Time to Collision (MTTC), Post-Encroachment Time (PET), and Deceleration Rate to Avoid Collision (DRAC), were used. The study employed extreme value models to estimate the risk of crashes based on the distribution of these conflict indicators. The crash risk was determined by calculating the probability of at least one indicator value exceeding its corresponding boundary value. The data for the study were collected from video footage and crash data obtained from four signalized intersections in Surrey, British Columbia. Various univariate, bivariate, and trivariate Bayesian hierarchical Extreme Value Models were developed and estimated using the data. The best-fitted models were selected based on their goodness-of-fit to statistical criteria. The results showed that traffic volume, shock wave area, and platoon ratio significantly influenced the crash risk. The estimated crash rates from the models were compared with the observed crash rates, and the models provided reasonable estimates within the 95% confidence intervals. Overall, the study demonstrated the applicability of the proposed crash estimation method using traffic conflict indicators and highlighted the importance of considering multiple indicators in crash risk assessment.

Ali et al. (2023b) proposed a real-time crash risk framework aimed at enhancing pedestrian safety at signalized intersections. By utilizing a Bayesian generalised extreme value model, they were able to establish a strong correlation between observed and predicted crashes.

Furthermore, the study involved the generation of separate generalised extreme value distributions, allowing for insights into both risky and safe signal cycles.

A study conducted by Alozi and Hussein (2022) took a different approach by leveraging two autonomous vehicle datasets collected from different locations in the United States and Singapore. Their focus was on modelling autonomous-pedestrian interactions through the development of a peak-over-threshold model. By incorporating various covariates such as turning movements and conflict speeds, the study estimated the expected number of collisions between autonomous vehicles and pedestrians, ranging from 2.3 to 5.5 per million vehicle kilometres travelled.

Similarly, a recent study by Arun et al. (2023) explored leading pedestrian intervals using a Bayesian quantile regression analysis. By estimating conflict thresholds and incorporating them into a Bayesian peak-over-threshold model, the researchers were able to assess the effectiveness of leading pedestrian intervals and their impact on pedestrian safety.

2.3.Conflict data collection methods

Data collection is a crucial aspect of conflict studies, and various techniques have been employed to gather traffic conflict information in the past decade. These techniques may vary vastly from one another and include examples such as driving simulators, video analytics, numerical simulations, connected vehicles, and autonomous vehicle probe data. Each method offers unique advantages and has been explained through representative studies described below.

Driving simulators provide controlled environments for collecting high-quality trajectory data and detailed driver demographics. For instance, Tarko (2012) proposes a new approach to modelling traffic interactions and crash causality. The study utilized a driving simulator experiment to estimate a peak-over-threshold model for road departures and near-departures, using data from four participants. The conflict severity was measured using a continuous measure of interaction severity. The proposed approach used the minimum time-to-collision (MTTC) as a measure of collision proximity and introduced the notion of collision proximity. The study suggested using various measures, such as post-encroachment time and minimum Euclidian distance, to calculate the interaction severity. The approach also incorporated an exposure-based model that related the frequency of crashes to the frequency of all vehicle interactions. The study used the Generalized Pareto distribution to model the tail behaviour of the severity distribution and estimated the frequency of crashes. The study

contributed towards a better understanding of the causality of the crash by appropriately capturing associated variables. The study emphasized the need to consider a continuum of traffic events and develop a plausible mechanism that connects different types of events to better understand crash occurrence. As seen with the above example, datasets from these simulator-based studies are often based on a limited sample size with an unrepresentative sample population causing unintended biases. Such a limited sample size in simulator studies is a common constraint that can impact final model performance.

Video-based conflict analysis studies analyse videos captured from intersections or freeway segments to extract meaningful data for model development. (Ali et al., 2023b) proposed a real-time crash risk estimation framework for vehicle-pedestrian interactions at signalized intersections. Unlike previous studies that commonly rely on static data sources, this study collected uninterrupted video recordings of vehicle movements at intersections and processed them using an AI-based video analytics platform. The extracted trajectory data and conflicts from the videos were fused with loop detector data containing information about traffic signal timing. The framework utilized an Extreme Value Theory approach to estimate crash risk by extrapolating frequently observed events (traffic conflicts) to rare events (crashes). The study addressed three key challenges in modelling traffic conflicts using Extreme Value Theory: small sample size, time-varying unobserved heterogeneity and capturing crash risk variation across different signal cycles and sites. A Bayesian Generalized Extreme Value model was developed, and the model parameters were estimated using a Bayesian estimation procedure. The study evaluated the model using the Deviance Information Criterion (DIC) and compared multiple models to select the best one based on local and global goodness-of-fit measures. The proposed framework was tested on three signalized intersections in Queensland, Australia. The study also collected crash data for benchmarking the models. The used video data processed through an AI platform to estimate generalised extreme value models for pedestrian safety demonstrated a close match with observed crashes. Covariates were extracted from the video footage and loop detector data to be included in the Extreme Value models. The study provides insights into the correlation between different covariates. Overall, the framework offers a proactive approach to estimating crash risk in real-time based on conflict indicators. Despite the high cost associated with data collection and processing, this method is frequently employed in the literature. However, this data collection method's time, labour and cost-intensive nature has limited the past studies to only a handful of locations.

Numerical simulations, conducted using microsimulation tools like AIMSUN/VISSIM, allow for the design of specific road facilities/interactions to assess safety. Wang et al. (2018) aimed to evaluate intersection safety by combining microscopic traffic simulation and Extreme Value Theory. Ten urban signalized intersections in Shanghai were selected, and field data, including traffic volumes and crash records, were collected. Simulation models were developed using a commercial microscopic simulation package (VISSIM) and calibrated using a measure of effectiveness and safety. Extreme Value Theory was used to develop empirical annualized crash frequency based on simulated conflicts and field conflicts. The results showed that the simulation-based empirical annualized crash frequency using the full-calibration strategy, performed the best in estimating crash frequency, especially for crossing and rear-end conflicts. The field-based empirical annualized crash frequency also showed good performance for rear-end conflicts but was not as effective for crossing and lane change conflicts. The time-to-collision-based measures performed poorly compared to the empirical annualized crash frequency. Overall, the study demonstrated the feasibility of using simulation-based empirical annualized crash frequency derived from Extreme Value Theory for intersection safety evaluation. However, most importantly, microsimulation lacks the human factor, which significantly influences driving behaviour and safety (Sharma et al., 2018).

LiDAR-based conflict analysis studies analyse point cloud data captured from intersections or freeway segments to extract meaningful data for model development. For example, (Wu et al., 2018) focused on the processing of roadside LiDAR data for the extraction of vehicle and pedestrian trajectories, as well as the identification of near-crash events between vehicles and pedestrians. The VLP-16 LiDAR sensor by Velodyne LiDAR™ was used for data collection, which generates a 360° 3D point cloud. Various algorithms were developed to perform background filtering, lane identification, object clustering, object classification, and data association. The trajectories of road users were successfully obtained, and the accuracy of object classification exceeded 93%. For near-crash identification, a novel method was introduced that considers the time difference to the point of intersection, the distance between stopped vehicles and pedestrians, and the speed-distance profile of vehicles. This method provided a systemic approach to identifying vehicle-pedestrian near-crashes. The study demonstrated the feasibility and effectiveness of using roadside LiDAR data to analyse road user behaviour and enhance safety analysis at intersections.

Autonomous vehicles equipped with advanced sensors have the capability to capture detailed information about their surroundings. Hu et al. (2022) processed Waymo data, recently

made publicly available for research purposes by Waymo (Sun et al., 2020), and demonstrated its application in analysing traffic safety. One advantage of autonomous vehicle data is the potential for network-level analysis, as these vehicles act as probes, collecting data for all road users in their catchment zone. However, there is limited evidence of applying autonomous vehicle data for corridor-level pedestrian safety.

2.4.Autonomous vehicle sensor data-based pedestrian safety studies

Various past studies have utilised autonomous vehicle sensor data to analyse pedestrian safety, providing valuable insights into the interactions between autonomous vehicles and pedestrians. Some notable research papers in this domain are described below.

Kutela et al. (2022) aimed to gain a deeper understanding of the patterns and underlying factors associated with autonomous vehicle (AV) crashes involving Vulnerable Road Users (VRUs). They employed text network analysis to map and analyse the narratives of 252 AV crashes involving VRUs, including pedestrians, bicyclists, and electric scooter users. The findings revealed that bicyclists and electric scooter users were more frequently involved in AV crashes directly, while pedestrians were predominantly involved indirectly. Direct AV crashes involving VRUs often occurred when the AVs were in autonomous mode, with the rear-left side of the AV being the most commonly affected. On the other hand, indirectly involved crashes resulted in more significant damages, particularly to the rear bumper of the AV. To predict VRU-involved crashes, the researchers utilized four classifiers, with the Random Forest classifier demonstrating the best performance. The important features for classification encompassed variables such as crosswalks, streets, and specific vehicle actions like turning and stopping. Similarly, Liu et al. (2021) conducted studies using data from the California Department of Motor Vehicles to identify and analyse autonomous vehicle-pedestrian conflicts. Although these studies did not specifically develop extreme value theory models, they provided important insights into the determinants of such conflicts.

Beauchamp et al. (2022) focused on data collected from dedicated autonomous shuttles in Quebec, Canada. The study aimed to assess the safety of automated vehicles by comparing their interactions with other road users to those of human drivers. The methodology involved collecting video data of AV shuttles operating in Quebec, Canada. User trajectories were extracted and classified into pedestrians, cyclists, motorized vehicles, and AV shuttles. Trajectories were then clustered to predict road user motion and identify control vehicles for the comparison. Interactions between AV shuttles and other road users were analysed using

various safety indicators, including speed, acceleration, time-to-collision (TTC), and post-encroachment time (PET). Statistical models were used to explore the associations between variables and safety indicators. They developed extreme value models for autonomous vehicle-pedestrian interactions, considering conflict measures such as time-to-collision, post-encroachment time, headway, and speed difference. The study found that autonomous shuttles exhibited safer behaviour than conventional vehicles, with lower operational speeds and acceleration. The findings provided insights into AVs' safety performance and interactions with different types of road users, helping inform future AV development and deployment efforts.

Alozi and Hussein (2022) employed empirical autonomous vehicle datasets from Lyft and nuScenes to investigate autonomous vehicle-pedestrian interactions. The study applied an Extreme Value Theory to analyse two autonomous vehicle datasets and predict pedestrian collisions. The datasets included trajectory data, LiDAR point clouds, and annotated video data collected from AV fleets operated by Motional and Lyft. The datasets were pre-processed to extract AV-pedestrian conflicts and calculate two conflict indicators, namely Post-Encroachment Time (PET) and Time-to-Collision (TTC). A total of 726 AV-pedestrian conflicts were identified, and the means and standard deviations of the conflict indicators were reported. The study also analysed a manually driven vehicle (MDV) dataset to validate the accuracy of the predicted collisions. The results showed that the Extreme Value Theory models could reasonably estimate the number of MDV-pedestrian collisions at signalized intersections. Covariates such as vehicle speed and pedestrian speed improved the accuracy of the models. The same Extreme Value Theory methodology was then applied to the AV data, and the expected number of AV-pedestrian collisions per million vehicle kilometres travelled (VKT) was estimated. The results indicated higher collision rates compared to MDV-pedestrian collisions, which can be attributed to the urban environment and the abundance of crosswalks and intersections. The study estimated a range of 2.3 to 5.5 autonomous vehicle-pedestrian collisions per million vehicle kilometres, depending on the specific model used. The study discussed the limitations of using trajectory data and highlighted the value of the Extreme Value Theory approach as a proactive method for predicting collisions.

However, the application of autonomous vehicle sensor data to assess pedestrian safety at a corridor level is an area that remains relatively unexplored. While these studies have shed light on the interactions between autonomous vehicles and pedestrians, further research is needed to fully leverage the potential of autonomous vehicle data for corridor-level pedestrian safety analysis.

2.5.Traffic conflict measures for pedestrian safety

Assessing pedestrian safety using traffic conflicts necessitates selecting suitable measures that effectively capture vehicle-pedestrian interactions (Arun et al., 2021b). It is crucial to select appropriate, theoretically justified conflict measures for assessing the specific conflict. Researchers must consider the specific research objectives, the nature of the interactions being examined, and the desired level of detail in the analysis. Past researchers have employed a range of conflict measures to examine various aspects, such as the complexity of evasive manoeuvres, object proximity, and crash severity (Kaparias et al., 2010; Laureshyn et al., 2010).

Some of the notable examples of conflict indicators used in the literature are listed below in Table 2.1. The table also briefly summarises the definition of the conflict measure along with some example studies in which those conflict indicators were predominantly used. For a detailed review of traffic conflict measures, refer to past studies that have extensively covered the topic (Ali et al., 2023a; Arun et al., 2021b; Arun et al., 2021c).

Table 2.1. Traffic conflict indicators and their definitions.

Conflict Measure	Definition	Example Study
Post-Encroachment Time (PET)	The time gap between when a conflicting vehicle exits the potential collision zone and when the subject vehicle, which has the right of way, reaches the potential collision point.	Alozi and Hussein (2022)
Time to Collision (TTC)	Time to collision between two vehicles under consistent collision course and speed differential.	Sobhani et al. (2013)
Gap Time (Gap)	In a leading-following scenario, it refers to the duration between the rear of the leading vehicle crossing a point on the road and the front of the following vehicle.	Pawar and Patil (2017)
Modified Time to Collision (MTTC)	A modified version of TTC that estimates the time to collision based on the relative speed and relative acceleration of the interacting vehicles.	Essa and Sayed (2019)
Time Exposed TTC (TET)	Cumulative duration, over a specified time period, for which the time to collision (TTC) value between two interacting vehicles remains below the TTC threshold for conflicts.	Rahman et al. (2019)
Yaw Rate (YR)	The angular velocity of a road user's rotation around the z-axis or the rate at which the heading angle changes.	Tageldin et al. (2015)

Time to Avoid a Collision (TTAvoid)	The time interval between the detection of a potential collision and the moment just before the collision is avoided.	Li et al. (2018)
Time to Brake (TTB)	The time period preceding the need for full braking to avoid a collision.	Char and Serre (2020)
Safety Index (Saf. I)	An index that combines the probability and severity of conflicts, calculated by weighting the released kinetic energy of a hypothetical crash by the Post Encroachment Time (PET).	Ismail et al. (2011)
Time Headway (TH)	The time duration between the front or rear of the leading vehicle and the front or rear of the following vehicle crossing a specific point on the road.	Zhao et al. (2020)
Required Deceleration Rate (RDR)	The rate of deceleration required for a vehicle to safely stop, considering the time headway between the subject vehicle and the conflicting vehicle equal to the PET.	Guido et al. (2011)
Rear-end Crash Potential (RECP)	A temporal measure calculated based on the headway, the driver's perception-reaction time, and the time required for braking.	Weng and Meng (2014)
Lateral Acceleration / Deceleration (Lat A / Lat D)	The instantaneous acceleration or deceleration of a vehicle in the lateral direction of motion.	Guo et al. (2010)
Kinetic Energy Loss per Unit Mass (DKE)	The kinetic energy released as a result of a collision between vehicles of similar masses.	Ma et al. (2018)
Jerk	The rate at which acceleration changes over time.	Hu et al. (2022)
Encroachment Time (ET)	The duration in which an offending vehicle violates the right-of-way of another vehicle.	Zhang et al. (2012)

Among the diverse pool of conflict measures utilized in pedestrian safety studies, temporal proximity measures have emerged as the most commonly adopted in the literature (Ghadirzadeh et al., 2022; Santhosh et al., 2020). These measures include post-encroachment time, which assesses the duration between a pedestrian leaving the encroachment area and a vehicle subsequently entering the same area, and time-to-collision, which quantifies the

temporal distance to a potential collision between a vehicle and a pedestrian if their velocities remain unchanged.

It is crucial to select appropriate conflict measures that are theoretically justified for assessing pedestrian safety. Researchers must consider the specific research objectives, the nature of the interactions being examined, and the desired level of detail and accuracy in the analysis. The choice of conflict measures should align with the research framework and contribute to a comprehensive understanding of vehicle-pedestrian conflicts and their implications for pedestrian safety. In conclusion, researchers must carefully consider the theoretical justifications and applicability of these measures to ensure meaningful and accurate assessments of pedestrian safety.

2.6.Summary and research gaps

Based on a comprehensive review of the literature, the following key insights can be drawn regarding pedestrian safety. Firstly, studies in the field have commonly relied on crash-based statistical models, primarily focusing on both mid-blocks and signalized intersections. However, these models predominantly utilize police-reported data, which may suffer from limitations such as under-reporting and a lack of detailed behavioural information. Secondly, to address the limitations of police-reported data, researchers have increasingly turned to traffic conflict-based pedestrian safety assessment frameworks such as extreme value theory models. This approach provides an alternative perspective that offers valuable insights into pedestrian safety beyond traditional crash-based analysis. Thirdly, various methods have been employed for collecting data on traffic conflicts, with video data being the most prominently utilized, particularly at intersections. This approach allows for extracting meaningful information about conflicts for efficient extraction of road user trajectories and other behavioural information. Lastly, recent advancements in vehicle technologies, such as autonomous vehicles, present unprecedented opportunities to enhance our understanding of traffic safety, including pedestrian safety. While a few applications of autonomous vehicle data concerning pedestrian safety can be found in the literature, there is still a lack of comprehensive exploration and utilization of autonomous vehicle data for analysing pedestrian safety at a corridor level. This research gap serves as a key motivation for the present study, aiming to address this knowledge limitation and harness the potential of autonomous vehicle data for corridor-level pedestrian safety analysis.

Chapter 3 Methodology

The methodology employed in this study comprises several key steps to achieve the objectives and address the research questions laid out at the end of section 1. Figure 3.1 below illustrates the overall framework of the methodology adopted in this research.

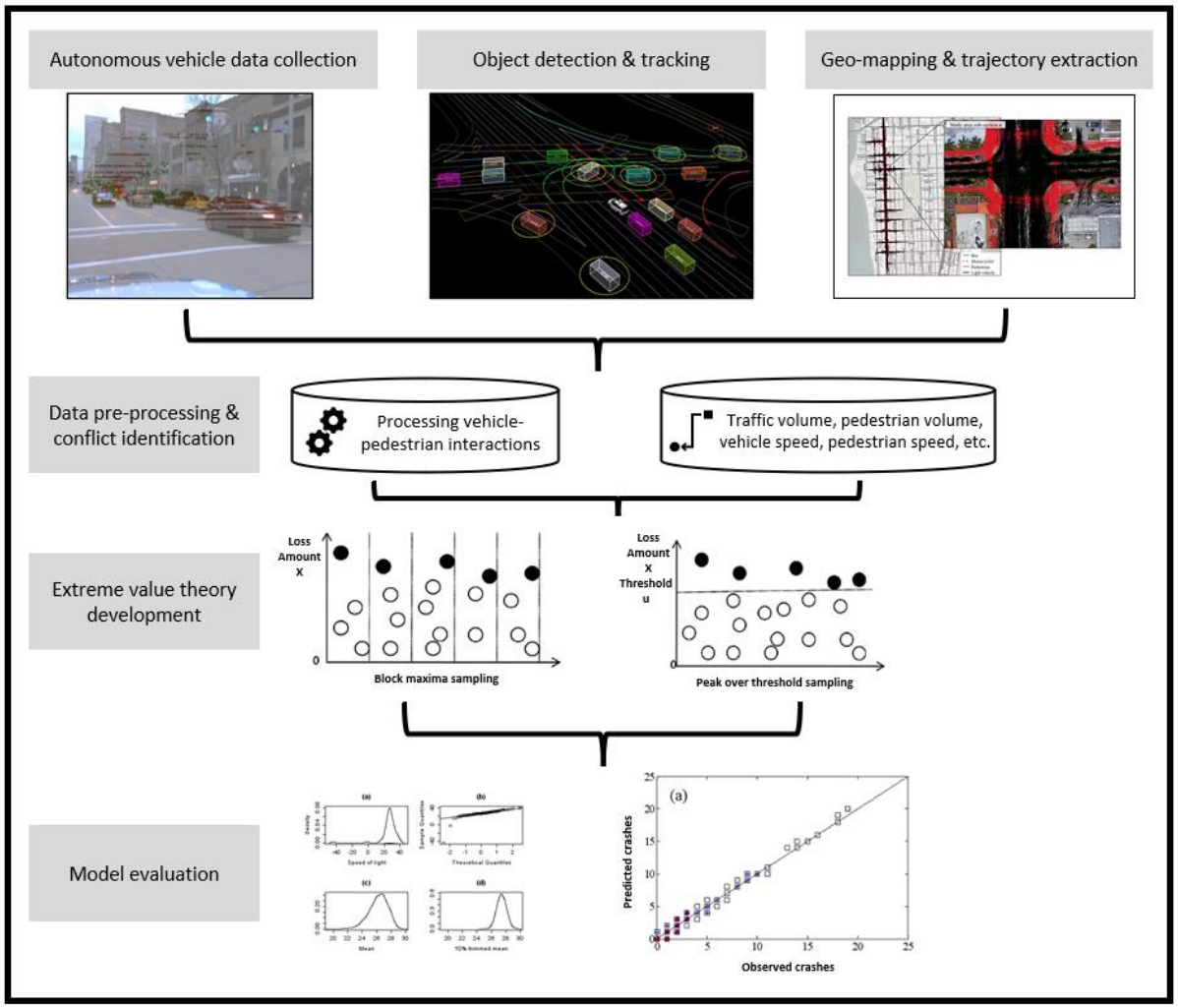


Figure. 3.1. The proposed pedestrian crash risk assessment framework using autonomous vehicle sensor data.

The first step involves the collection and pre-processing of autonomous vehicle data. This step includes activities such as blending multiple data files, ensuring data quality through thorough checks, georeferencing the data to establish spatial references, and applying filters to remove any irrelevant or noisy data.

Next, in the second step, the data is prepared for the application of extreme value theory modelling. This step includes extracting object trajectories from the autonomous vehicle data, analysing object trajectory pairs from each scenario to identify potential traffic conflicts, and

identifying relevant covariates information that may influence the occurrence of extreme events.

The third step centres around the development of the extreme value theory model. This step involves fitting the model to the input data, conducting tests to assess the influence of covariates on extreme events, and selecting the most appropriate model based on statistical criteria and local goodness-of-fit measures.

Finally, in the fourth step, the developed model is thoroughly evaluated. Global goodness-of-fit assessment based on real-life observed data from the study corridor is conducted to evaluate the model's overall performance.

3.1.Autonomous vehicle data and pre-processing

This section succinctly describes the steps involved in collating and pre-processing the Autonomous Vehicle sensor data as received from Argoverse. The collation and pre-processing steps generate output which is used for the extraction of conflict data along with other relevant information used in the extreme value model development process.

The first step involves creating a folder directory that maps all 250,000 unique episode IDs. This directory serves as a systematic reference index of the data, allowing easy access and retrieval of specific episodes for analysis. Next, the hierarchal structure of the Argoverse dataset is transformed into a tabular data frame format for the scenario data within each episode. This conversion enables a more structured representation of the data, facilitating efficient data manipulation and analysis. From the scenario data, all object trajectories are constructed using the available 110 frames in each episode. These trajectories capture the movement and paths of various objects, such as vehicles, pedestrians, and cyclists, throughout the duration of the scenario. To ensure compatibility and consistency between all the available episodes, the trajectories are converted from the local coordinate reference system to the widely used WGS84 reference system. This conversion aligns the data with the global coordinate system, allowing for seamless integration and comparison with other geospatial data sources. In order to filter and focus on objects within the study area, the converted trajectories are plotted on a geo-referenced map. This visualization helps identify and select the objects that fall within the specific geographical region of the study area. Further detailed information on data pre-processing can be found in Chapter 4.

3.2.Extreme Value Model input data processing

The autonomous vehicle data processing steps begin by loading the scenario file and map file that have been filtered to include only the episodes within the study area, as determined in the previous step. To ensure the quality and accuracy of object annotation, trajectory sanity checks are performed. This step involves cross-referencing the geographical and contextual information gathered from the combination of map data and scenario data. Any incorrect or erroneous object annotations are corrected during this step. Next, two data frames are created to separate the pedestrian and vehicle information extracted from each scenario. These data frames contain the trajectories of the identified pedestrians and vehicles, which will be used for further analysis. The pedestrian and vehicle trajectories are overlaid in pairs in order to identify potential conflicts. Next, the encroachment zone is defined using the pedestrian and vehicle object dimensions and velocity vectors. This encroachment zone helps determine the critical space where conflicts may occur. Following that, important parameters such as post-encroachment time, vehicle velocity, acceleration, and other supplementary information are calculated and stored. Post-encroachment time for two conflicting road users refers to the time difference between the instance the first road user exits the course of the second traffic participant, also known as the encroachment zone, and the following road user enters the same course as illustrated in Figure 3.2. These parameters provide additional insights into the dynamics of the interactions between pedestrians and vehicles. Covariates, which are additional factors that may influence the occurrence of conflicts, are extracted from the data. These covariates include the total pedestrian count, vehicle count, average pedestrian speed, average vehicle speed, and other relevant variables. These covariates are utilized in non-stationary extreme value theory models. Finally, all the processed data, including trajectories, encroachment information, covariates, and other relevant data, are exported to a CSV and shape file. This file serves as the input for fitting extreme value theory models based on the processed data.

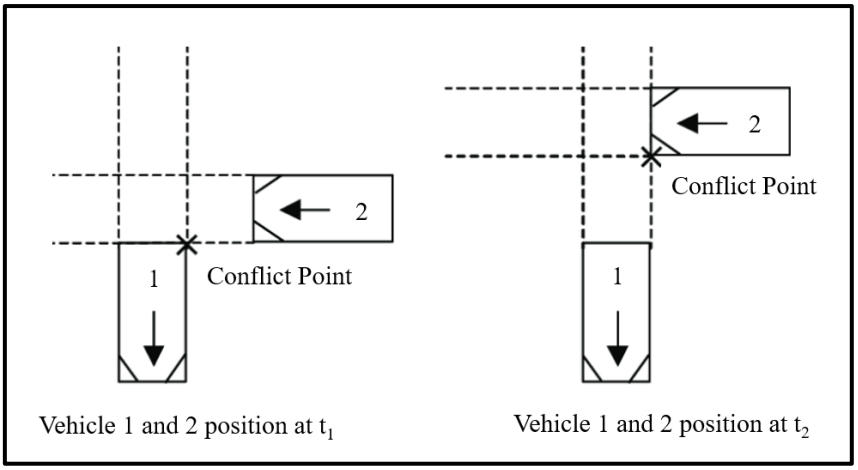


Figure. 3.2. Post-encroachment time illustration.

Further detailed discussion on the data processing step is provided in Chapter 4 after briefly introducing the dataset used in the study.

3.3.Extreme Value Model development

Extreme Value Model introduction

Extreme value modelling is a statistical technique for analysing and modelling rare and extreme events. These events, often referred to as "extremes", occur infrequently but have a significant impact when they do occur. Examples of such events include large floods, severe storms, rare disease outbreaks, and extreme financial market fluctuations. Understanding and predicting these extreme events is crucial for risk assessment, decision-making, and developing appropriate mitigation strategies.

The extreme value modelling technique focuses on analysing the tail-end characteristics of a distribution where extreme events are located. Traditional statistical methods, such as Gaussian or normal distribution models, are inadequate for accurately capturing extreme tails. On the other hand, extreme value modelling is specifically designed to model the distribution of extreme values, providing more reliable estimates and insights into extreme events. The field of extreme value modelling encompasses various statistical approaches to sample extreme events from a given dataset. The two most widely used approaches in the literature are block maxima and peak over threshold sampling approaches. Block maxima extreme sampling approach based on generalized extreme value distribution, Peak-over-threshold extreme sampling approach based on generalized Pareto distribution and their application in this study are described in further detail in the model development section.

Extreme value modelling finds applications across numerous disciplines, including finance, hydrology, climate science, engineering, and environmental risk assessment. In finance, for instance, extreme value models help estimate the risk associated with rare events and evaluate the tail risk of financial portfolios. In hydrology, these models are used to analyse and predict extreme river flows or rainfall intensities, aiding in the design of flood protection measures and water resource management strategies. In the realm of road safety, extreme events encompass severe and rare traffic incidents that have the potential to cause significant harm and damage. Traditional statistical methods, such as Gaussian or normal distribution models, fail to accurately capture the extreme tail behaviour where these events lie. Extreme value modelling, on the other hand, specifically addresses this limitation by concentrating on the analysis of extreme values. It offers statistical tools and techniques to estimate the probabilities, magnitudes, and frequencies of rare and severe traffic incidents, thus enhancing our understanding of road safety risks and providing a solid foundation for proactive safety analysis and intervention. This property makes extreme value theory a perfect mathematical tool to underpin the conflict-based pedestrian safety framework developed in this study.

The following sections will briefly overview the mathematical details of extreme value theory, the modelling methodology, and the covariates used in this research project.

Model development

At the heart of the modelling section of this study's framework lies the application of Extreme Value Theory, which enables the estimation of rare events, such as crashes, based on more frequently observed events, like traffic conflicts. To achieve this, it is crucial to sample extreme events from the available data, which forms a vital step in the modelling process. The literature commonly employs two sampling approaches: Block Maxima and Peak Over Threshold (Ali et al., 2023a). This study aims to utilize both of these modelling approaches and provides a detailed explanation of their implementation.

The Block Maxima approach involves sampling extreme event observations from fixed time or space blocks. Selecting the appropriate block interval is crucial for obtaining a well-fitted model. In video analytics-based studies, where observation periods can span from days to months, this approach is straightforward to apply by selecting different block sizes, such as 5 minutes, 10 minutes, or 20 minutes. However, when using autonomous vehicle data, a challenge arises due to the format of the empirical autonomous vehicle datasets, which provide short episodes of data rather than continuous streams. In this study, the Block Maxima

approach is applied at the episode level, with each episode lasting 11 seconds. The maximum value of the conflict indicator from each episode is considered an extreme event, and the negated post-encroachment time serves as a conflict indicator for modelling vehicle-pedestrian safety. To apply the Block Maxima approach, consider $x_1,\ x_2 \ldots\ x_n$ are a sequence of random and independent variables with a common distribution function, with $M_n = max[x_1,\ x_2 \ldots\ x_n]$ providing block maximum of *n* values, and will lead to generalised extreme value distribution when n $\rightarrow \infty$. Mathematically, the generalised extreme value distribution function can be expressed as

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3.1}$$

where μ, σ, and ξ are the location, scale, and shape parameters of the generalised extreme value distribution, respectively.

Another sampling approach is the event-based Peak Over Threshold method. In this approach, observations above a predetermined threshold are considered extreme events. Determining the threshold is vital to the model's performance as it affects the sample size and target curve used in the model fitting process. Setting a low threshold may result in an abundance of observations, violating the asymptotic assumption, while a very high threshold may lead to insufficient sample size for reliable model estimation. Threshold determination is typically carried out using mean residual life plots and threshold stability plots (Coles, 2001), which are further discussed in subsequent sections.

Succinctly, conflict extremes identified from a series of observations can be used to estimate a Peak Over Threshold sampling-based model. Assume that $x_1,\ x_2 \ldots\ x_n$ represents independent and identically distributed random observations, the cumulative distribution function of exceedances *X* over the threshold *u* can be obtained as

$$Fu\,(x) = P(X - u \leq x | X > u). \tag{3.2}$$

The distribution can be approximated as generalised Pareto distribution for a sufficiently high value of threshold *u* as

$$G(y) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi},\ \xi \neq 0 \tag{3.3}$$

where σ and ξ are the scale and shape parameters of Generalised Pareto distribution, respectively.

However, practical challenges needed to be addressed when applying extreme value theory models to this study, which will be discussed in detail in Chapter 5 under the section Model Development.

In order to apply extreme value theory models to the context of this study, some practical challenges need to be considered. Vehicle-pedestrian interactions are highly influenced by several determinants, such as traffic volume and pedestrian volume, and not accounting for the effects of those covariates can lead to time-varying unobserved heterogeneity issues, which is likely to affect model performance. As such, several covariates affecting vehicle-pedestrian interactions are incorporated into the model to handle the non-stationarity of traffic conflict extremes and capture unobserved heterogeneity. To this end, the Block Maxima and Peak Over Threshold sampling-based models are parameterised with the strict assumption that the scale parameter must be positive, $\emptyset = log\ \sigma$. If z_{ij} represents i^{th} episode maximum for episode j, the Generalised Extreme Value distribution and the Generalised Pareto distribution can be represented by Eq. (3.4) and Eq. (3.5), respectively.

$$G\left(z_{ij} < z|\mu_{ij}, \phi_{ij}, \xi_{ij}\right) = exp\left\{-\left[1 + \xi_{ij}\left(\frac{z-\mu_{ij}}{exp\ (\phi_{ij})}\right)\right]^{-1/\xi_{ij}}\right\}. \quad (3.4)$$

$$G\left(z_{ij} < z|\phi_{ij}, \xi_{ij}\right) = 1 - \left[1 + \left(\frac{\xi_{ij}\, z}{\exp(\phi_{ij})}\right)\right]^{-\frac{1}{\xi_{ij}}}. \quad (3.5)$$

Several covariates are included in modelling parameters using identity link functions as

$$\begin{pmatrix} \mu_{\mathrm{ij}} = \alpha_{\mu 0} + \alpha_{\mu 1}\boldsymbol{X} + \varepsilon_{\mu \mathrm{j}} \\ \phi_{\mathrm{ij}} = \alpha_{\phi 0} + \alpha_{\phi 1}\boldsymbol{Y} + \varepsilon_{\phi \mathrm{j}} \\ \xi_{\mathrm{ij}} = \alpha_{\xi 0} + \varepsilon_{\xi \mathrm{j}} \end{pmatrix}, \quad (3.6)$$

where, $\alpha_{\mu 0}$, $\alpha_{\phi 0}$, and $\alpha_{\xi 0}$ are model parameter intercept terms, $\alpha_{\mu 1}$ and $\alpha_{\phi 1}$ are parameter estimates for the covariate vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively and $\varepsilon_{\mu j}$, $\varepsilon_{\phi j}$, and $\varepsilon_{\xi j}$ are random error terms.

The Bayesian parameter estimation approach is adopted in this study, which mathematically captures observed data, abstraction of observed data, and the inherent uncertainty of model parameters (Smith, 2020). Further, the Bayesian model estimation procedure offers flexibility in estimating posterior distribution by specifying priors in

parameter estimation[1]. Due to no prior information on distribution parameters, normally distributed priors with zero mean and large variance are used. Markov Chain Monte Carlo simulation with Gibbs sampling technique is used to obtain the posterior distribution of model parameters.

Model covariates

Unobserved heterogeneity is a common problem that limits the performance of stationary extreme value models due to underlying mechanisms that drive traffic conflicts. Covariates play a crucial role in extreme value models, providing valuable insights into the factors that influence conflicts. These covariates capture the characteristics and dynamics of the system under study, enabling a deeper understanding of the underlying factors contributing to extreme outcomes. By incorporating covariates into the modelling process, it becomes possible to account for the effects of various factors on the occurrence and severity of rare events. Covariates can range from basic characteristics such as traffic volume and road geometry to more complex factors like weather conditions and driver behaviour. This study focused on four covariates in particular to capture the characteristics of extreme events.

The "vehicle volume in an episode" covariate quantifies the total number of vehicles present in the conflict episode within a specified distance from the encroachment zone. This information reflects the level of traffic and potential interactions among vehicles near the conflict area.

The "pedestrian volume in an episode" covariate represents the total number of pedestrians present in the episode within the defined proximity to the encroachment zone. This covariate helps capture the presence and volume of pedestrians in the vicinity of the conflict, which is an important factor influencing the likelihood of vehicle-pedestrian interactions.

The "average vehicle speed" covariate calculates the average speed of all moving vehicles in the episode. This covariate provides insights into the general speed behaviour of vehicles in the vicinity of the conflict, which can influence the severity and outcome of potential conflict.

Similarly, the "average pedestrian speed" covariate computes the average speed of all moving pedestrians in the episode. This covariate captures the typical walking speed of

[1] Bayesian extreme value modelling technique incorporated processed conflict data from the entire corridor, thereby resolving data scarcity and non-stationarity problem.

pedestrians near the conflict area, which can affect the time available for vehicles to respond and avoid collisions.

3.4.Model performance evaluation

The methodology for evaluating model performance involves both local and global assessments. Locally, the deviance information criterion (DIC) is utilized to measure the model's goodness of fit. This criterion compares the model's complexity and fits the observed data, providing insight into its local performance.

Local model performance

Several models are estimated and compared in this study, and to determine the best model, Deviance Information Criterion is used as a model goodness-of-fit. Mathematically,

$$DIC = \overline{\mathrm{D}} + p_d, \tag{3.7}$$

where $\overline{\mathrm{D}}$ and p_d are posterior mean deviation and the effective number of model parameters, respectively. The model with the least deviance information criterion is preferred over its competing models.

3.5.Model validation

Globally, the model is evaluated by comparing its predictions against observed crashes and their corresponding confidence intervals. This comprehensive approach allows for a thorough evaluation of the model's performance at both the local and global levels, ensuring a robust assessment of its predictive capabilities.

Global model performance

The developed models are used to estimate crashes for a specified period, which are compared with historical crash records. Specifically, the mean crash estimates and confidence intervals of the estimated crashes are compared to those of observed crashes, whereby the mean crashes can be computed as

$$N = \frac{\tilde{T}}{T} RC, \tag{3.8}$$

where N is the expected number of crashes for the duration $\tilde{T}$, T is the observational period, and RC denotes the risk of a crash. To understand the uncertainty associated with crash estimates and compare that to the observed one, the confidence intervals are calculated. For the observed crashes, the Poisson confidence interval for the true mean is estimated as

$$\frac{1}{2y}\chi^2_{2n,(1-\alpha/2)} < \lambda < \frac{1}{2y}\chi^2_{(2n+1),\alpha/2}, \quad (3.9)$$

Where y is the number of years of observation, *n* is the number of observed events, χ^2 is the chi-square critical value and α is the significance level. However, the confidence interval for model crash estimates is obtained using a simulation process (Songchitruksa and Tarko, 2006). As the model estimations are a scalar function of the parameters and are assumed to follow normal distribution under regularity conditions, confidence intervals can be obtained from the quantiles of the empirical distributions obtained from the simulation process. One hundred thousand simulations are set to run, and upper and lower bounds based on a 95% confidence interval are obtained.

Chapter 4 Autonomous vehicle dataset

The advent of autonomous vehicle technology has brought about significant advancements in transportation research and development. One of the key resources associated with Autonomous vehicles is the extensive amount of sensor data they collect during their operation. These datasets can provide a valuable means for researchers and practitioners to gain insights into various aspects of transportation, including safety analysis, mobility patterns, and infrastructure planning. In recent years, several autonomous vehicle companies and research initiatives have made efforts to share their collected data with the wider community, leading to the availability of publicly accessible datasets.

One prominent example of publicly available autonomous vehicle sensor datasets is Waymo's Open Dataset. Waymo, a leading autonomous vehicle company, has released a substantial collection of sensor data captured from their self-driving vehicles. This dataset encompasses diverse driving scenarios, including urban environments, highways, and challenging weather conditions. It includes high-resolution sensor data such as lidar point clouds, camera images, and radar data, enabling researchers to analyse and develop advanced algorithms for perception, localization, and mapping.

Another significant source of autonomous vehicle sensor data is the Lyft Level 5 autonomous vehicle dataset. Lyft, a well-known ride-sharing platform, has created a dataset containing sensor data collected from their autonomous vehicles. This dataset includes lidar, camera, and other sensor data captured during real-world driving operations. The dataset covers a range of urban driving scenarios and offers a valuable resource for researchers working on perception and navigation algorithms, as well as for evaluating the performance of autonomous vehicle systems.

Argoverse is yet another noteworthy initiative in the field of autonomous vehicle datasets. Argoverse has released a large-scale dataset comprising high-definition maps, sensor data, and vehicle trajectories collected from their autonomous vehicle research fleet. This dataset focuses on complex urban environments and provides rich information for tasks such as motion forecasting, scene understanding, and behaviour prediction.

These publicly available autonomous vehicle sensor datasets offer researchers and practitioners the opportunity to explore real-world autonomous vehicle data, which was previously limited in accessibility. They allow for the development and evaluation of advanced

algorithms and methodologies, contributing to the progress of autonomous vehicle technology and its applications in various domains, including transportation safety, urban planning, and intelligent transportation systems.

The following section will briefly describe the datasets that were explored for this study, along with a quick explanation of factors that played a critical role in the selection of the final dataset used in the study. The methodology used to process the data to make it fit for conflict-based modelling purposes is also described in the following sections.

4.1.Publicly available datasets

Waymo dataset

The Waymo dataset includes high-definition object trajectories produced using an onboard perception system alongside stationary and non-stationary map features that offer perspective for the road environment. Object trajectories are sampled at a rate of 10Hz and contain information such as the object's bounding box (including 3D centre point, heading, length, width, and height) and velocity vector. However, due to sensor limitations or occlusions, there are some missing measurements for some time steps, which are indicated by a valid flag. The map data consists of polylines and polygons representing lane centres, speed bumps, lane boundaries, stop signs, road edges, crosswalks, and traffic signals, including the lanes they control. The map features also store specific data related to different feature types, such as the type of lane boundary (e.g., broken white or double yellow).

In the dataset, 20-second segments are compiled from various road user interactions. These segments are further divided into 9.1-second scenes consisting of 91 steps at 10Hz. The data is then split into a 70% training set, a 15% validation set, and a 15% test set. Two versions of the validation and test sets are available in the dataset: the standard and interactive versions. Both versions focus on the 9.1-second scene with a focus on different objects. For research requiring longer time frames, the original 20-second scenarios are also made available.

An illustration of the object information available from scenes in the dataset is provided in Figure 4.1 below.

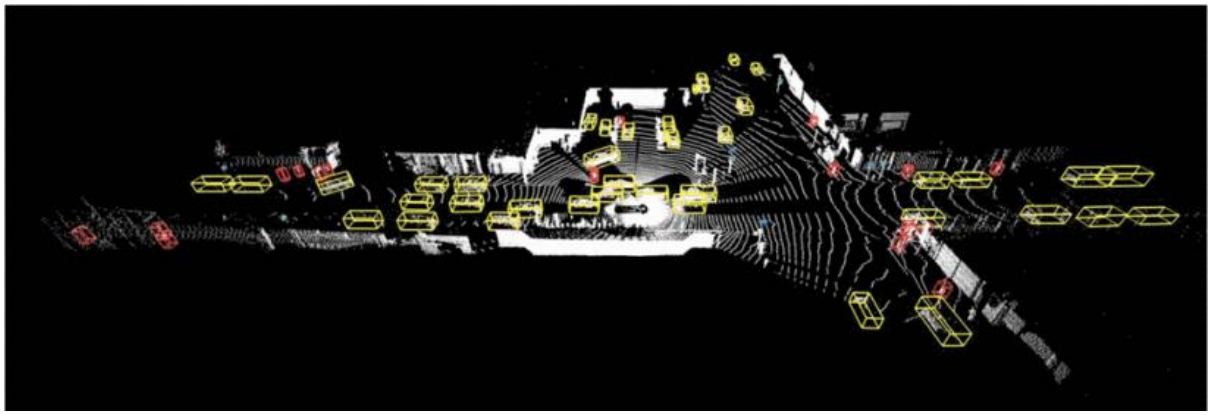


Figure. 4.1. Waymo Dataset example episode [Copyright © 2020, IEEE].

Argoverse 2 motion dataset

The Argoverse 2 Motion Forecasting dataset provides a collection of prediction episodes gathered from a self-driving fleet. The dataset comprises 250,000 non-overlapping episodes, randomly split into 80% for training, 10% for validation, and 10% for testing. These episodes are mined from six distinct urban driving environments in the United States. The dataset includes a total of 10 object types, with five falling into the dynamic category (e.g., vehicles, pedestrians) and five into the static category (e.g., buildings, road signs).

Each episode in the dataset consists of a local vector map and 11 seconds of trajectory data (captured at a frequency of 10 Hz) for all tracks observed by the ego-vehicle within the local environment. In each episode, a single track is designated as the "focal agent," ensuring that its observations are complete throughout the entire duration of the episode. The selection of the focal agent aims to maximize interesting interactions with map features and nearby actors.

Additionally, a subset of tracks is identified as "scored actors" to evaluate multi-agent forecasting. These actors are carefully chosen to guarantee episode relevance and minimum data quality threshold.

An illustration of the object information available from scenes in the dataset is provided in Figure 4.2 below.

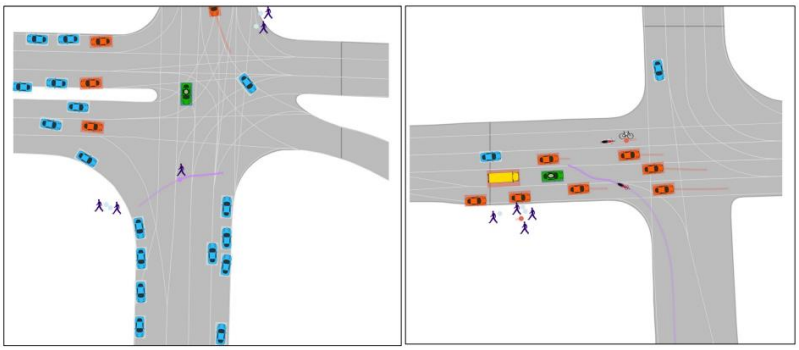


Figure. 4.2. Argoverse Dataset example episode [Adapted from (Wilson et al., 2023)].

Lyft dataset

The dataset encompasses a collection of 170,000 scenes, each lasting 25 seconds, resulting in a cumulative log duration of over 1,118 hours. These scenes were recorded by a fleet of self-driving vehicles following a predetermined route. The dataset includes data from various perception sensors, including seven cameras, 3 LiDARs, and five radars.

Specifically, one LiDAR is positioned on the vehicle's roof, while two LiDARs are mounted on the front bumper. The roof-mounted LiDAR features 64 channels and rotates at a frequency of 10 Hz, whereas the bumper-mounted LiDARs have 40 channels. The seven cameras, situated on the roof, collectively offer a 360° horizontal field of view. Additionally, four radars are installed on the roof, with one radar placed on the forward-facing front bumper.

Data collection for this dataset took place between October 2019 and March 2020, specifically during daytime hours between 8 AM and 4 PM. For each scene, the observable road users were detected, including pedestrians, vehicles, and cyclists.

Each road user is internally assigned a 2.5D cuboid, along with information such as velocity, acceleration, yaw, yaw rate, and class label. Lyft’s proprietary perception system was employed to detect these traffic participants. The system fuses data from multiple sensor modalities to provide a comprehensive 360° view of the surrounding environment for self-driving vehicles.

The dataset was split into training, test, and validation sets, employing an 83-7-10% ratio. Each self-driving vehicle contributes data exclusively to one set. The dataset is encoded in n-dimensional compressed zarr arrays.

An illustration of the object information available from scenes in the dataset is provided in Figure 4.3 below.

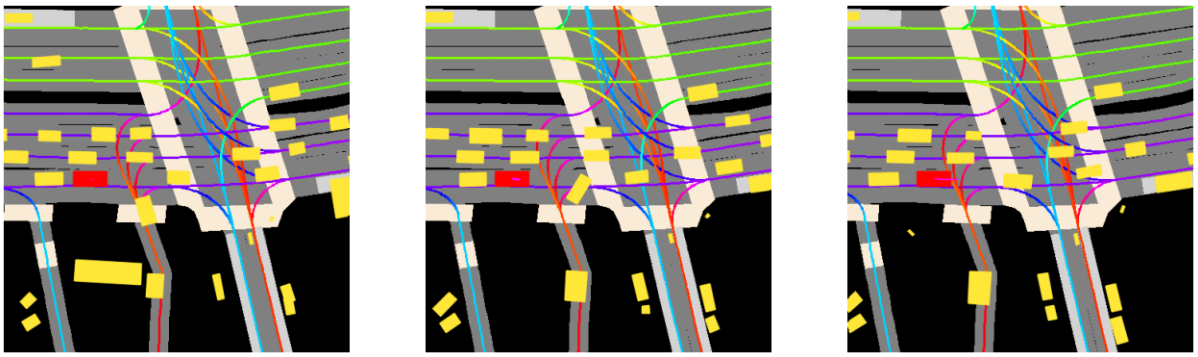


Figure. 4.3. Lyft Dataset example episode [Adapted from (Houston et al., 2020)].

Dataset comparison

Several empirical autonomous vehicle datasets have been developed to facilitate research and development in the field of autonomous driving. Prominent datasets include Lyft Level 5 autonomous vehicle dataset, Waymo Open Dataset, Argoverse Dataset, NuScenes Dataset, and Yandex Self-Driving Car Dataset. These datasets vary in terms of scale, sensor modalities, data collection environments, and annotation details. The Lyft dataset provides high-quality sensor data captured by a fleet of AVs, while the Waymo dataset offers diverse sensor data and detailed annotations. The Argoverse dataset focuses on urban episodes with high-definition maps, and the NuScenes dataset covers a wide range of urban driving environments. The Yandex dataset provides data from a Russian perspective, capturing unique driving conditions.

Table 4.1 below summarises key comparisons for a few datasets from a growing list of options available in the space of autonomous vehicle research.

Table 4.1. Key autonomous vehicle dataset comparison.

Measure	**Lyft**	**Waymo**	**NuScene**	**Yandex**	**Argoverse 2**
No. of Episodes	170k	104k	41k	600k	250k
Total Time	1118h	574h	5.5h	1667h	763h
Episode Duration	25s	9.1s	8s	10s	11s
Sampling Rate	10Hz	10Hz	2Hz	5Hz	10Hz
No of Cities	1	6	2	6	6
Unique Roadways	10km	1750km	-	-	2220km
No. of Object Class	3	3	1	2	5

The following section will highlight the dataset selected for this research and the reasons behind the choice. It will discuss the dataset layout and structure in further detail.

4.2.Dataset selected for the research

The study utilises the publicly available Argoverse 2 dataset as a source of autonomous vehicle sensor data. Argoverse conducted autonomous vehicle trials across six cities in the United States and released multiple datasets from these trials. Encompassing a significant geographic

area across six diverse cities, the Argoverse 2 dataset spans over 2,000+ km. It features a comprehensive object taxonomy with ten distinct classes covering both static and dynamic actors. Specifically, the study focuses on the Argoverse 2 Motion Forecasting Dataset, which comprises a curated collection of 250,000 episodes. Each episode spans a duration of 11 seconds and provides the 2D, birds-eye-view centre point and trajectory information of each tracked object, collected at a frequency of 10 Hz. The curation process involves a thorough analysis of a massive amount of driving data obtained from the fleet of autonomous test vehicles with a focus on identifying segments that present the most challenging and atypical behaviour, particularly from road users relevant to the driving decision selection process of the autonomous vehicle. The dataset captures a wide range of interactions, including complex manoeuvres such as vehicles yielding to pedestrians at crosswalks, buses at multi-lane intersections, and cyclists navigating through dense city streets. The diversity and long duration of episodes incentivize the development of methods that excel in ensuring safety in challenging and rare situations. Researchers and developers can leverage this dataset to advance the performance and robustness of motion forecasting algorithms for autonomous vehicles, addressing real-world complexities and promoting the development of a safer road network.

Why Argoverse?

This section will delve into the factors that guided the decision-making process when selecting a particular autonomous vehicle dataset over alternative options. The key factors such as dataset size, geographic coverage, and object diversity which formed the criteria to choose the dataset most suitable for the research objectives, are outlined below.

Expansive roadway coverage

The dataset generated by the Argo fleet covers an extensive roadway network, which spans a total of more than 2200 km. This wide coverage ensures that the dataset captures a diverse range of driving environments, including various road types, traffic conditions, and geographical regions. The extensive coverage enhances the dataset's representativeness.

Diverse road user annotation

The dataset maps and annotates five road user classes, including buses and bicycles. This rich representation of various road user classes allows for multi-user safety analysis, enabling researchers to examine interactions, behaviours, and potential conflicts between different types of road users.

Embedded geo-information

The third reason is the embedded geo-information within the dataset. The dataset provides geo-referenced data, allowing for easy comparison and integration with other data sources. The inclusion of geo-information facilitates the alignment of the dataset with external geographic data, such as maps or satellite imagery, enabling a more comprehensive analysis.

Argoverse autonomous vehicle sensor setup

The data was collected using a fleet of hybrid vehicles that were fully integrated with Argo's artificial intelligence self-driving technology, classified as Level 4 according to the Society of Automotive Engineers (SAE). SAE states that L4 vehicle can perform all driving tasks and monitor the environment without human intervention, but only within specific operational design domains (ODD). Level 4 vehicles can operate autonomously in certain conditions or environments, such as a specific city or highway. Argoverse did not comment on the connectivity level of their autonomous vehicles, hence the vehicles are assumed not to be connected/communicating to each other. The vehicles, as shown in Figure 4.4, were equipped with two roof-mounted LiDAR sensors (model number VLP-32C) with 64 beams in total, offering an unobstructed range of 200 meters. These LiDAR sensors generated an average point cloud of approximately 107,000 points per second at a frequency of 10 Hz. Additionally, the fleet had seven high-resolution cameras and two front-facing stereo cameras recording at a frequency of 20 Hz, providing a combined 360° field of view. The sensor setup used in the Argoverse fleet is depicted in the Figure 4.4 below.

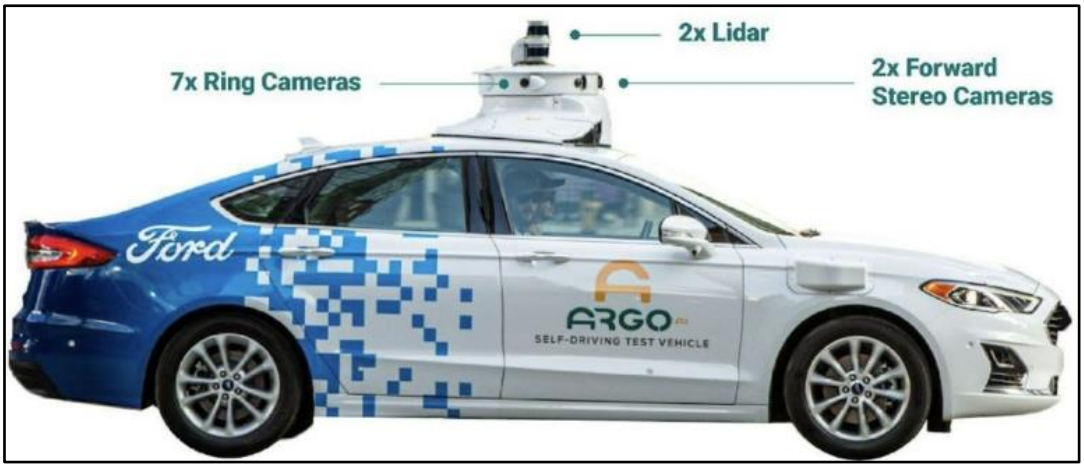
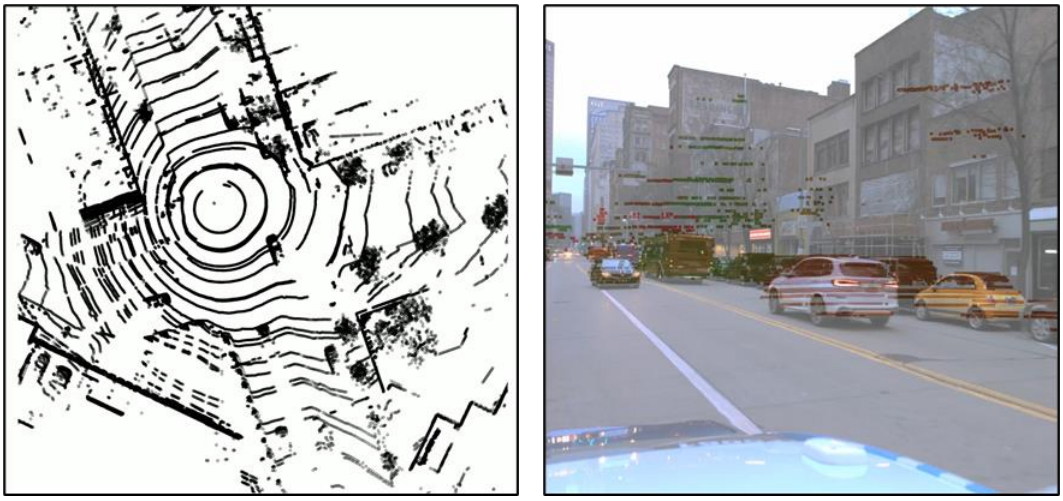


Figure. 4.4. Typical Argoverse autonomous vehicle [Adapted from (Wilson et al., 2021)].

Figure 4.5 below shows an example of Argo autonomous vehicle fleet vehicle manoeuvring on a public road in Texas, US. The LiDAR sensors, the stereo cameras and the ring cameras generate synchronous data. The left-side figure depicts a bird’s eye view of the

point cloud generated by LiDAR sensors mounted on the roof of the vehicle. The right-side figure depicts the rear-left ring camera feed overlayed with non-stationary points from the LiDAR point-could. This information is then processed through a self-supervised machine learning algorithm by Argoverse to annotate different objects detected and their trajectories recorded along the route.



(a) LiDAR output (b) Camera output

Figure. 4.5. Example of sensor output from Argoverse [Adapted from (Wilson et al., 2021)].

Argoverse data format

The Argoverse dataset is encoded in a proprietary hierarchical format, and the study employed Agro's open-source API to decode the data into a structured tabular format. Each decoded episode contains object-related information and map-context information, which are merged together to obtain detailed data for analysis.

The map file is provided in a json hierarchical structure. The file contains map log id unique to each episode and individual classified object vector boundaries for drivable areas, pedestrian crossings, lane boundaries for each lane and lane markings type information. It also contains relational information on the geospatial relation between lanes such as left neighbour, right neighbour, predecessor, successor, etc. All this information is stored in a city-specific geo-coordinate system, which can be converted to various global coordinate reference systems using a transformation matrix provided by Argoverse. The content of a typical map file from an episode is illustrated in Figure 4.6 below.

```
static_map

ArgoverseStaticMap(log_id='0008c251-e9b0-4708-b762-b15cb6effc27', vector_drivable_areas={13204115: DrivableArea(id=13204115, ar
ea_boundary=[Point(x=2099.99, y=753.94, z=-23.86), Point(x=2100.31, y=752.44, z=-23.85), Point(x=2100.95, y=750.0, z=-23.93), P
oint(x=2082.84, y=750.0, z=-23.82), Point(x=2081.88, y=753.74, z=-23.85), Point(x=2077.95, y=768.95, z=-23.54), Point(x=2075.2
1, y=779.4, z=-23.32), Point(x=2070.47, y=798.02, z=-23.05), Point(x=2069.04, y=803.87, z=-22.98), Point(x=2067.4, y=810.0, z=-
22.83), Point(x=2085.52, y=810.0, z=-23.05), Point(x=2085.71, y=809.22, z=-23.08), Point(x=2087.65, y=801.76, z=-23.17), Point
(x=2090.42, y=790.75, z=-23.33), Point(x=2090.6, y=790.35, z=-23.34), Point(x=2090.8, y=790.08, z=-23.34), Point(x=2091.22, y=7
89.69, z=-23.33), Point(x=2091.47, y=789.55, z=-23.31), Point(x=2091.78, y=789.43, z=-23.28), Point(x=2092.52, y=789.36, z=-23.
23), Point(x=2093.33, y=789.48, z=-23.19), Point(x=2095.35, y=790.06, z=-23.18), Point(x=2096.66, y=783.01, z=-23.25), Point(x=
2096.22, y=782.95, z=-23.25), Point(x=2094.34, y=782.55, z=-23.29), Point(x=2093.89, y=782.3, z=-23.33), Point(x=2093.54, y=78
1.95, z=-23.4), Point(x=2093.32, y=781.68, z=-23.44), Point(x=2093.17, y=781.4, z=-23.48), Point(x=2093.06, y=781.03, z=-23.5
1), Point(x=2093.05, y=780.7, z=-23.52), Point(x=2093.58, y=778.55, z=-23.52), Point(x=2096.11, y=768.7, z=-23.69), Point(x=209
7.51, y=763.2, z=-23.78), Point(x=2097.83, y=762.75, z=-23.78), Point(x=2098.24, y=762.43, z=-23.75), Point(x=2098.57, y=762.3
2, z=-23.71), Point(x=2098.91, y=762.24, z=-23.7), Point(x=2099.49, y=762.26, z=-23.64), Point(x=2101.51, y=762.73, z=-23.52),
```

Figure. 4.6. Static map information from a typical episode.

The scenario file is provided in a paraquet hierarchical structure. It contains a unique scenario id for each episode or snippet consistent with the map log id in the map file. This enables the matching and mapping of two files to generate road user trajectories. The scenario file contains detailed nanoseconds timestamps of each frame captured by the sensors. For all the objects identified and labelled in a scenario, individual tracks with unique track id and object location, velocity, acceleration and heading details are included. The content of a typical scenario file from an episode is illustrated in Figure 4.7 below.

```
scenario

ArgoverseScenario(scenario_id='0008c251-e9b0-4708-b762-b15cb6effc27', timestamps_ns=array([3.15981382e+17, 3.15981382e+17, 3.15
981382e+17, 3.15981382e+17,
       3.15981382e+17, 3.15981382e+17, 3.15981382e+17, 3.15981382e+17,
       3.15981382e+17, 3.15981382e+17, 3.15981383e+17, 3.15981383e+17,
       3.15981383e+17, 3.15981383e+17, 3.15981383e+17, 3.15981383e+17,
       3.15981383e+17, 3.15981383e+17, 3.15981383e+17, 3.15981383e+17,
       3.15981384e+17, 3.15981384e+17, 3.15981384e+17, 3.15981384e+17,
       3.15981384e+17, 3.15981384e+17, 3.15981384e+17, 3.15981384e+17,
       3.15981391e+17, 3.15981391e+17, 3.15981391e+17, 3.15981391e+17,
       3.15981391e+17, 3.15981391e+17, 3.15981391e+17, 3.15981391e+17,
       3.15981392e+17, 3.15981392e+17, 3.15981392e+17, 3.15981392e+17,
       3.15981392e+17, 3.15981392e+17, 3.15981392e+17, 3.15981392e+17,
       3.15981392e+17, 3.15981392e+17]), tracks=[Track(track_id='133051', object_states=[ObjectState(observed=True, timestep=0,
position=(2079.1146118846427, 952.765728969816), heading=3.1216457341451584, velocity=(-0.050521500921899484, -0.00508642157198
4299)), ObjectState(observed=True, timestep=1, position=(2079.0994930634224, 952.7676009573411), heading=3.12128015490937, velo
city=(-0.03594407845472953, -0.003945697139784356)), ObjectState(observed=True, timestep=2, position=(2079.1051444829764, 952.7
690674429542), heading=3.120977402440795, velocity=(-0.040044767937584 08, -0.0029295827704551173)), ObjectState(observed=True,
```

Figure. 4.7. Scenario object information from a typical episode.

All 250,000 episodes in the database were processed to extract starting position of Autonomous vehicles for all six cities. Figure 4.8 geographically illustrates all the scenarios extracted from the database.

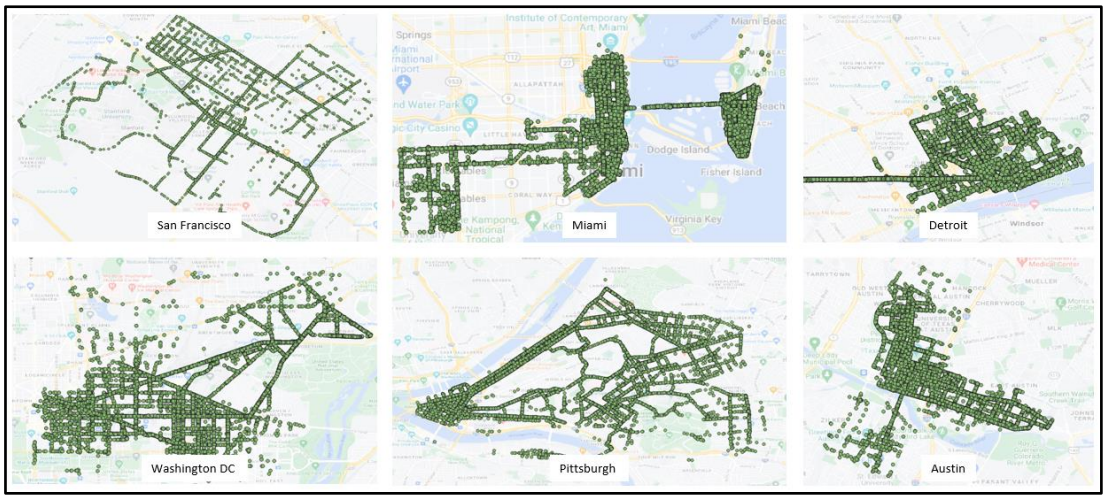


Figure. 4.8. Argoverse dataset area coverage across six cities in the US.

Further analysis of the number of episodes and the geographical coverage of the fleet movement in the cities helped shortlist the study area. Table 4.2 below summarises the episode count and area coverage for all six cities. Miami was shortlisted as it had the highest count of episodes and the widest area coverage.

Table 4.2. Argoverse Motion Forecast dataset properties by city.

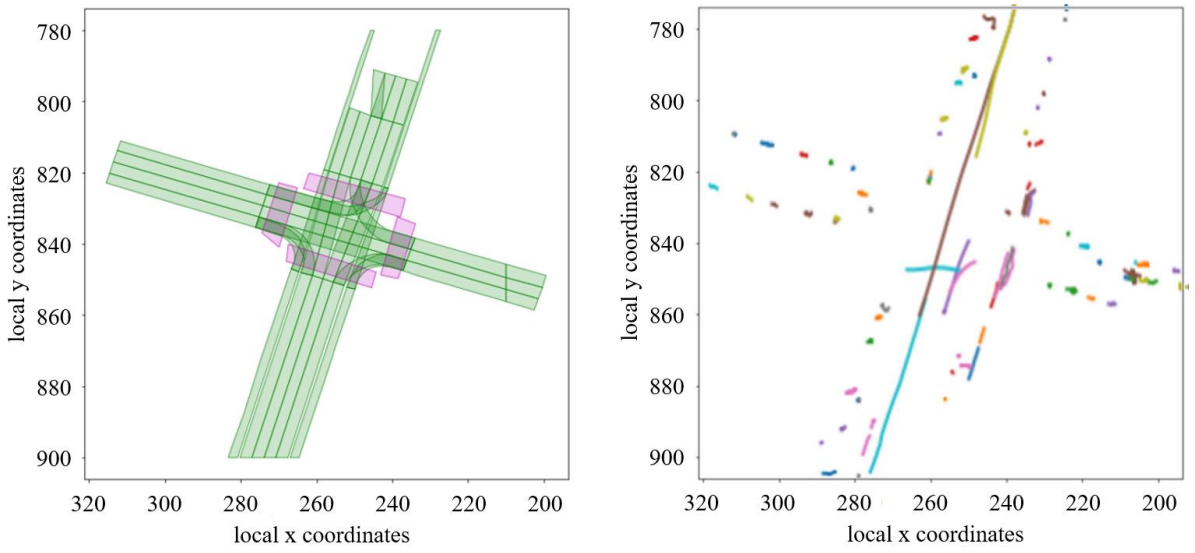
City	Total Episodes	Total Area Covered (km^2)
Austin	52919	27
Detroit	30084	37
Miami	67358	44
Pittsburgh	53305	25
San Francisco	14594	25
Washington DC	31642	27

4.3.Data processing methodology

The data processing methodology is set up to effectively handle the autonomous vehicle sensor data from Argoverse, called data pre-processing here onwards, and extract the relevant information from the dataset to estimate extreme value models, called data processing here onwards.

The data pre-processing stage involves several steps to combine and analyse the Argoverse dataset that provides extensive geographical coverage and a substantial amount of travel data. These pre-processing steps ensure the data is organized, standardized, and ready for subsequent analysis. The following steps are undertaken to transform the data for the research project:

1) Episode Mapping: In the process of working with autonomous vehicle data, one of the initial steps is episode mapping. This involves creating a directory that maps all the unique episode IDs present in the dataset. Organizing the episodes in a structured manner makes it easier to access and analyse specific episodes during the subsequent data processing and modelling stages.
2) Scenario and Map File Decoding: The scenario and map files need to be decoded to extract relevant information from autonomous vehicle datasets. Decoding involves converting the proprietary hierarchical format of the data into a more accessible and standardized tabular structure. Decoding the files improves the interpretability and further processing time for the data. Figure 4.9 below demonstrates a visual example of one episode's map and scenario files.



(a) Episode map file (b) Episode scenario file

Figure. 4.9. Argoverse 2 dataset files visual representation.

3) Data Collation and Visualization Generation: After decoding the scenario and map files, each episode's data must be collated and combined into a comprehensive dataset. This step involves merging the relevant information from the scenario and map files. Additionally, birds' eye view motion trajectory visualizations are generated to provide a better understanding of the data and aid in identifying object motion in each episode.
4) Trajectory Construction: One of the key components of autonomous vehicle data analysis is constructing object trajectories. This step involves extracting and tracking the movement of objects, such as vehicles or pedestrians, throughout the duration of an

episode. To do so, the position coordinates of all the objects are tracked throughout an episode and joined together to construct the trajectories.

5) Coordinate Reference System Conversion: Autonomous vehicle data comes in a local coordinate reference system specific to each individual city. To facilitate comparisons and integration with other geospatial datasets, the dataset is converted into a common coordinate reference system, WGS84. This conversion ensures consistency and enables spatial analysis and visualization across different datasets, such as crash datasets. Figure 4.10 represents all the objects detected from the dataset from Miami City plotted over a geo-referenced map.

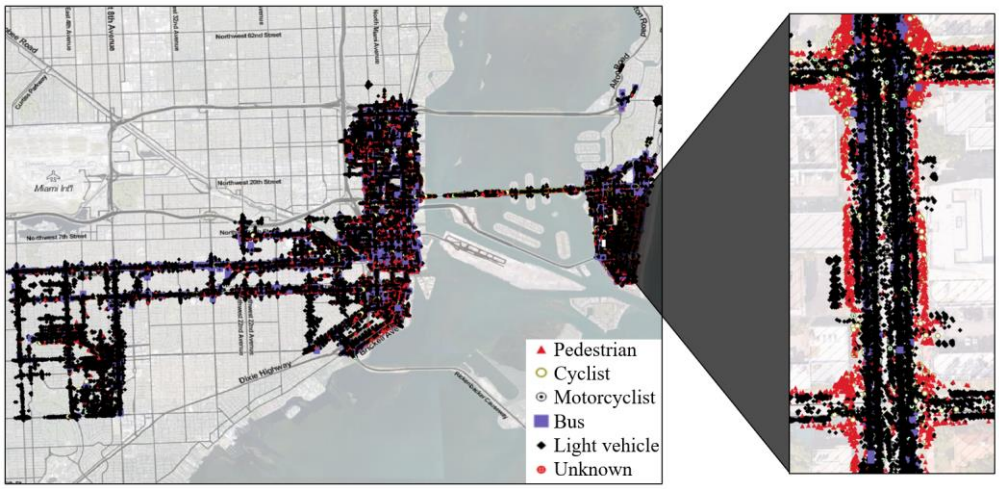


Figure. 4.10. Starting position of all objects in the Miami city dataset.

6) Study Area Selection: To define the scope and focus of the investigation, a corridor was selected to form the focus of the study. The study area city selection was based on the maximum number of varied traffic interaction episodes. Miami, Florida, was selected as the city of focus as it provided the maximum number of total episodes with the widest area coverage. As the study primarily focuses on pedestrian safety, all the object trajectories from over 67 thousand episodes in Miami were plotted on a geo-referenced map to select a study corridor with rich pedestrian activity, as seen in Figure 4.11. This led to the selection of Alton Road between 6th Street and 17th Street in Miami Beach (see the green area in Figure 4.11). This 19 km corridor comprised 15 intersections and 14 mid-blocks classified into 15 sub-sections. A total of 6,533 episodes were filtered from the Miami dataset for corridor analysis.

Figure. 4.11. Study corridor with a zoomed example of one sub-section.

After selecting the study area, data processing steps are implemented to analyse the pre-processed data. This process involves steps such as loading the data, analysing object trajectories, estimating conflict indicators, and extracting covariates. These data processing steps lay the foundation for subsequent modelling and analysis, allowing for a comprehensive understanding of the pedestrian crash risk within the selected study area. The following steps are undertaken in data processing for the research project:

1) Study Area Scenario and Map Loading: Once all the episodes in the dataset were tagged in or out of the study area, map and scenario files from the relevant episodes were loaded for further analysis. From over 67,000 episodes, comprising more than 3,300,000 individual object trajectories, a total of 6,533 episodes were loaded for corridor analysis consisting of more than 410,000 object trajectories. A typical example can be seen in Figure 4.12, which illustrates study area trajectories with a zoomed-in example of one sub-section.

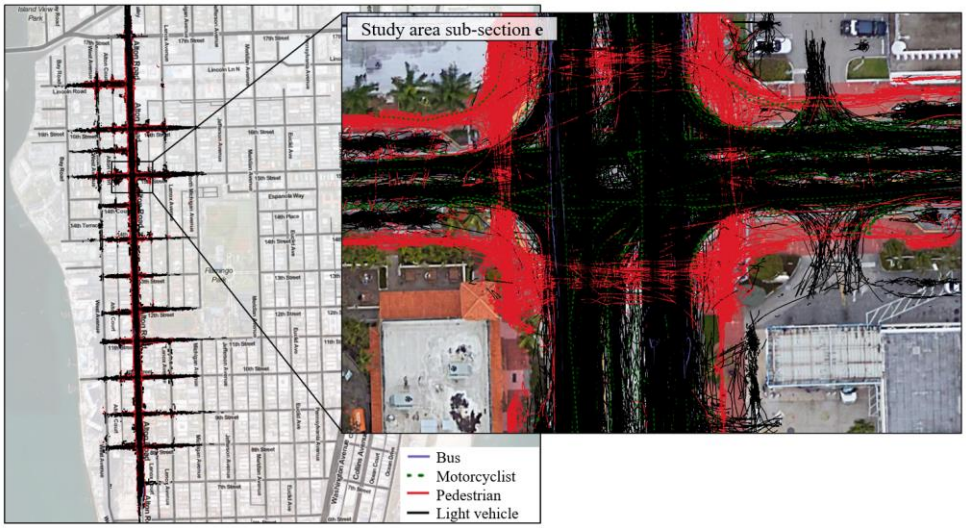


Figure. 4.12. Object trajectories extracted with a zoomed example of one sub-section.

2) Trajectory Analysis: The analysis of the dataset highlighted inconsistencies between object labels and their corresponding trajectories. Vehicles engaged in roadside parking manoeuvres and complex movements within mixed-use areas like parking lots and driveways generated unrealistic trajectories. For instance, the trajectories of drivers or passengers exiting parked vehicles often overlapped with the vehicle's trajectory. To address these issues and eliminate inconsistencies and noise, driveable areas and pedestrian crossing boundaries were extracted from map layers, which provided detailed information about driving lanes and designated pedestrian areas. This information was integrated into our algorithm for conducting sanity checks to ensure accurate and reliable object labels.
3) Object Class-based DataFrame Definition: Object class-based data frames are created to facilitate further analysis. This step involves segregating the data broadly into two classes pedestrians and other vehicles. Each object class contains the list of the objects identified in an episode belonging to that class and relevant attributes such as trajectories. All object trajectories from the two class objects are compared with each other to identify potential conflicts. For example, Figure 4.13 plots a potentially conflicting object pair of a pedestrian (highlighted in red) and a car (highlighted in blue) from one of the episodes.

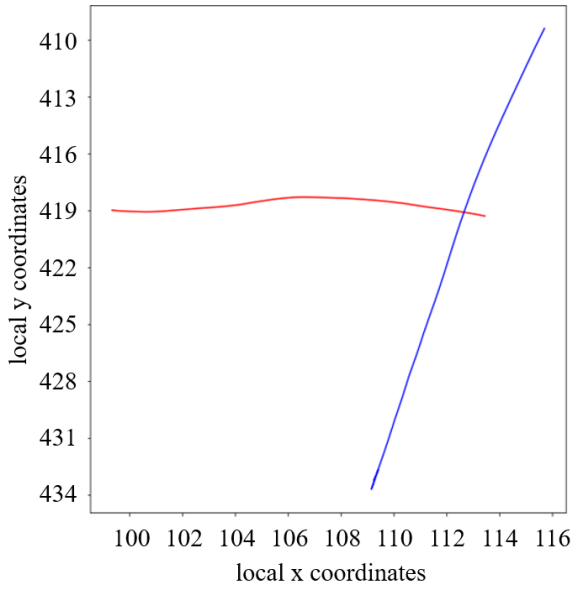


Figure. 4.13. Intersecting trajectory pair identification.

4) Encroachment Area Definition: The encroachment area refers to the zone within which a conflict or potential collision occurs. It is defined based on various factors, such as object dimensions, velocity, and trajectories. As Argoverse only provides centroid information for all the objects identified, standard dimensions for different object classes such as bus, car, and pedestrian were assumed. The encroachment area was used to calculate the post-encroachment time in the next step, which was used as the conflict indicator in the study.
5) Conflict Indicator Extraction: The post-encroachment time conflict indicator is extracted from the data to quantify and measure conflicts. Post-encroachment time refers to the duration between the moment two objects (e.g., a vehicle and a pedestrian) cross their encroachment zone. To determine the post-encroachment time, the positions, dimensions, and velocities of the objects are considered. From Figure 3.2 above in the previous section, post-encroachment time = $t_2 - t_1$
6) Covariate Extraction: Covariates are additional factors or variables that may influence or contribute to conflicts or safety risks. These covariates are extracted from the scenario data, such as the number of vehicles and pedestrians present in an episode and the average speeds of vehicles and pedestrians in an episode. By including covariates in the analysis, the accuracy and reliability of the models was enhanced.
7) Data Export for Modelling: After extracting the necessary information, the data is exported in a suitable format for modelling purposes. This step typically involves

exporting the data into a structured file, a CSV (comma-separated values) format, that can be easily imported into modelling software.

The final processed dataset from the study area had a total of 581 pedestrian-vehicle conflicts from 474 unique episodes. At each episode level, Table 4.3 shows the list of covariates that were extracted and used as input in the model for handling non-stationary.

Table 4.3. Statistical summary for conflict indicator and traffic flow variables.

Parameter	Mean	Standard deviation	Minimum	Maximum
Post encroachment time (s)	3.6	2.01	1.2	6
Vehicle volume in the episode	19.8	6.43	5	41
Pedestrian volume in the episode	3.1	2.36	1	13
Vehicle speed (m/s)	5.49	3.22	0	17.38
Pedestrian speed (m/s)	1.39	0.92	0	6.91

4.4.Crash dataset

The Florida Department of Transportation Crash Data Dashboard is a publicly available resource that visually represents general crash statistics in Florida. It presents the data in the form of graphs and charts, providing an overview of various types of traffic crashes reported by law enforcement. The dashboard encompasses all public roadways in Florida and allows users to filter the data based on criteria such as year, county, or pre-selected crash types. The information is refreshed monthly, but it can take law enforcement agencies up to 90 days to report crashes. The data source for the dashboard is the crash forms submitted by law enforcement agencies, and it is maintained by FLHSMV. The dashboard provides official, finalized data and statistics for all reported traffic crashes, making it a valuable tool for visualization and analysis. The data encompassed detailed information about each reported crash, including the location, time, weather conditions, types of vehicles involved, road user details, collision type, number of injuries or deaths, intoxication details, lighting, and severity of injuries. The FLHSMV website interface and its content, when filtered for pedestrian crashes between 2014-2020, is illustrated in Figure 4.14 below.

Figure. 4.14. Pedestrian crashes in Miami (2014-2020).

Pedestrian crash data for the study area were obtained from the Florida Department of Transportation (FLHSMV, 2022) to validate the developed extreme value theory models. Seven-year crash data (2014-2020) was extracted to provide valuable insights into traffic crashes and their characteristics around the study area. To accurately compare model results, the crashes within the study area involving pedestrians were filtered from the dataset.

Chapter 5 Results and discussion

5.1.Local model performance

The local model performance section, as described earlier in methodology section 3.4, provides an evaluation of the model's performance at a local level. Several models are estimated and compared in this study, and this section presents the results obtained through the application of the deviance information criterion as a measure of model fit. The deviance information criterion allows us to assess the goodness of fit of the model by considering both its complexity and how well it explains the observed data. The section also discusses the interpretation of the results obtained from the comparison of the local performance of various models developed in the study.

Block Maxima sampling-based model

Several Bayesian Block Maxima sampling-based models are developed and estimated in the Bayesian framework. Past studies used 50,000-100,000 simulation iterations with two chains for model convergence and posterior distribution estimation (Ali et al., 2022b; Kamel et al., 2022). Using the upper limit from the examples above, this study ran two chains with 100,000 iterations, whereby the first 50,000 were discarded as burn-in, and the remaining were used to calculate the posterior distributions of the model parameters. The convergence of the model was assessed using two diagnostics. First, a visual inspection of trace plots indicates that the chains are well-mixed, as illustrated in Figure 5.1. Second, the Gelman-Rubin statistic value for each parameter was calculated and found to be less than 1.1, reflecting the model convergence.

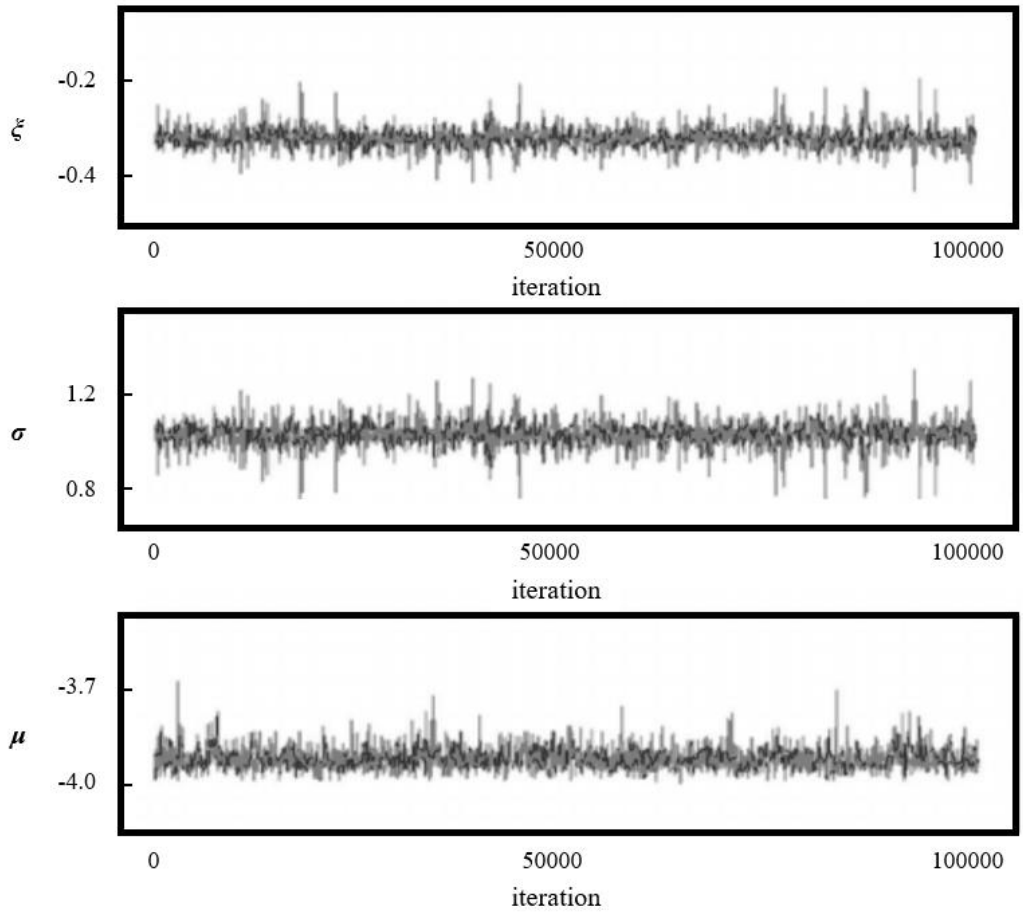


Figure. 5.1. An example trace plot for visual inspection.

Table 5.1 presents the deviance information criterion values used for comparing the three generalised extreme value models estimated in this study. All non-stationary models with location/scale parameters give a better fit (lower deviance information criterion values) than the stationarity model. Incorporating covariates in model estimation captures the variation in the data better, provides more insights into vehicle-pedestrian interactions, and thereby improves model goodness-of-fit. Comparing three possibilities of incorporating covariates (i.e., location, scale, and combined), covariates for a model with only scale parameter parameterisation were statistically insignificant and hence the model is not reported in Table 5.1. Out of the other two possibilities, the model with covariates incorporated into the location parameter possesses the lowest deviance information criterion value and is thus selected in the study.

The covariates incorporated in the analysis are pedestrian volume, vehicle volume, average pedestrian speed and average vehicle speed. The sign and magnitude of mean estimate of covariates in Table 5.1 can be used to interpret their impact on overall crash risk. Negative

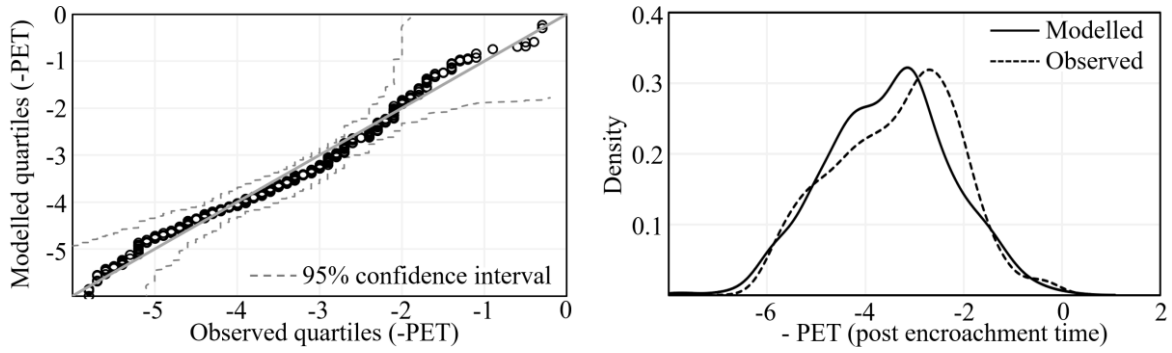
sign indicates inverse correlation with crash risk and greater magnitude indicates greater impact on overall crash risk. Further detailed discussion on selected models and their covariates can be found in Discussion section below.

Table 5.1. Summary of the Generalised Extreme Value model estimation results.

Model	Parameter	Location					Scale		Shape	DIC
		μ_0	μ_{PC}	μ_{VC}	μ_{PS}	μ_{VS}	$\emptyset_0$	$\emptyset_{VC}$	ξ_0	
Stationary	mean	-3.803	—	—	—	—	0.102	—	-0.295	
	s.d.	0.149	—	—	—	—	0.036	—	0.059	396
	2.5%	-3.950	—	—	—	—	0.059	—	-0.350	7
	97.5%	-3.658	—	—	—	—	0.103	—	-0.233	
Location parametrisation	mean	-3.074	-0.098	-0.022	-0.076	0.175	0.081	—	-0.276	
	s.d.	0.443	0.062	0.022	0.042	0.138	0.036	—	0.054	385
	2.5%	-2.653	-0.037	-0.000	-0.034	0.31	0.116	—	-0.222	0
	97.5%	-3.522	-0.158	-0.043	-0.116	0.039	0.046	—	-0.327	
Scale and location parameterisation	mean	-3.389	-0.101	—	-0.07	0.17	0.366	-0.014	-0.282	
	s.d.	0.404	0.059	—	0.044	0.137	0.182	0.012	0.055	386
	2.5%	-2.991	-0.043	—	-0.027	0.307	0.534	-0.002	-0.332	0
	97.5%	-3.783	-0.16	—	-0.112	0.037	0.178	-0.025	-0.223	

Abbreviations: PC = pedestrian volume; VC = vehicle volume; PS = average pedestrian speed; VS = average vehicle speed; DIC = Deviance Information Criterion; s.d. = standard deviation.

Figure 5.2 shows the goodness-of-fit of the selected model, which was used to further assess the model performance by performing a visual inspection of the probability density function of the empirical and modelled negated post encroachment time. These plots indicate that the model is reasonably well-fitted to the observed data because (a) the observations are found to lie along the line of equality (Figure 5.2 (a)) and (b) the modelled and observed curves (Figure 5.2 (b)) are very close to each other.



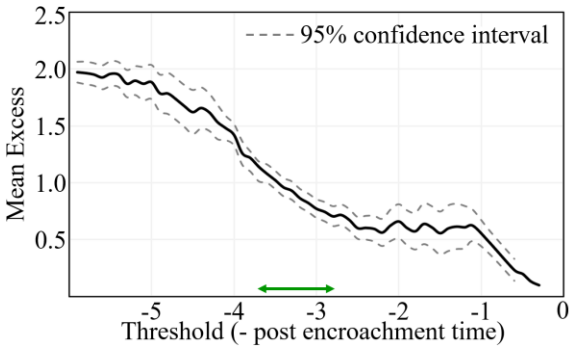
(a) modelled versus observed quartile plot (b) probability density plot

Figure. 5.2. Generalised extreme value model goodness-of-fit diagnostics.

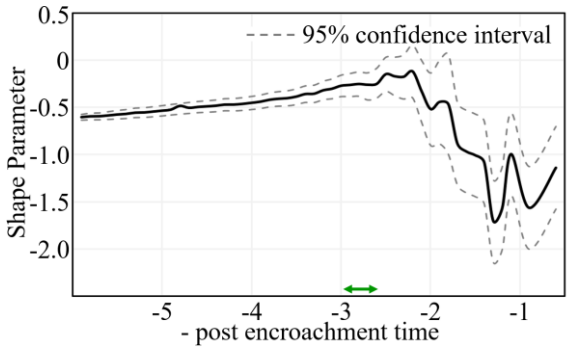
For the Block Maxima sampling-based models, the model with covariates in the location parameter outperforms other competing models in terms of both local and global goodness-of-fit measures. In a similar comparison of lane-changing crash risk, Ali et al. (2022a) concluded that the model with covariates in the location parameter performed better than other models.

Peak Over Threshold sampling-based model

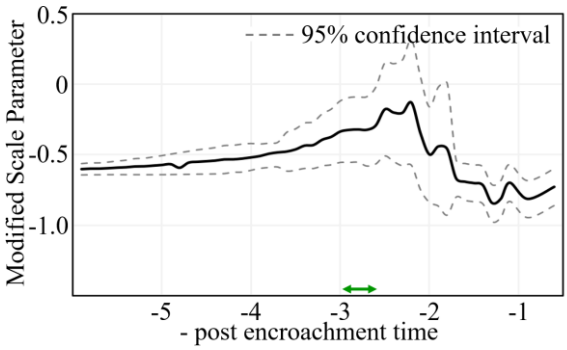
A dataset de-clustering process is applied for the Peak Over Threshold sampling-based model to account for any serial dependence. Pedestrian-vehicle conflicts involving common road users are identified and classified as clusters of dependent events, and only the highest extreme from the clusters is selected. An integral component of the model is the threshold, which needs to be appropriately determined. Following Coles (2001), the threshold is obtained using two plots: mean residual plot and modified scale and shape parameter plots. The mean residual life plot (Figure 5.3 (a)) exhibits linear behaviour in the sections between -3.7 and -2.8. However, the threshold needs to be based on the combination of all three plots. The shape and scale plots are constant between -3.0 and -2.6. Thus, the overlapping region between -3.0 s and -2.8 s is the ideal threshold range. The threshold is selected as the maximum value of the intersection of the three ranges, which is -2.8 s for this study, yielding 175 exceedance values.



(a) Mean residual Plot



(b) Shape parameter stability plot



(c) Modified scale parameter stability plot

Figure. 5.3. Diagnostic plots for the Peak Over Threshold sampling-based model.

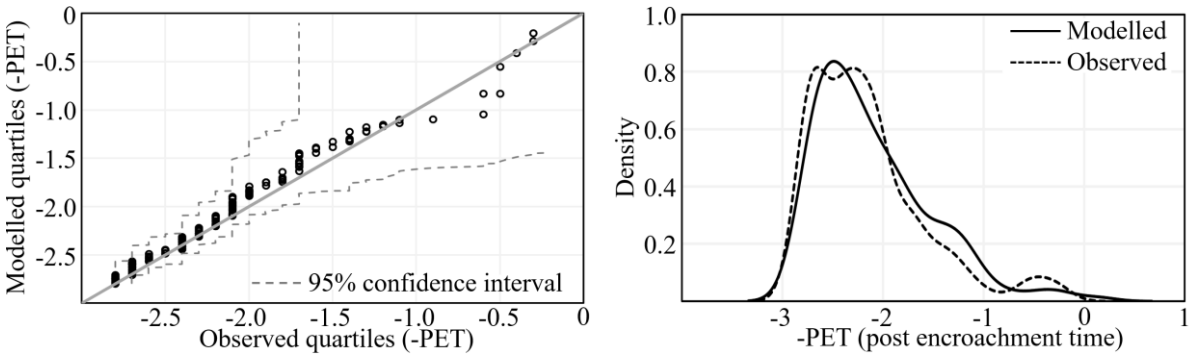
Table 5.2 presents the deviance information criterion values used for comparing the four Peak Over Threshold sampling-based models estimated in this study. The non-stationary models contain pedestrian and vehicle volume as covariates, whereas covariates regarding their speed were found to be insignificant and thus omitted from the parsimonious model. Akin to the Block Maxima sampling-based model, all non-stationary models reveal a better fit (lower deviance information criterion values) than the stationarity model. The model with covariates related to the volume of pedestrians and vehicles shows the best fit out of all the competing models based on the lowest lower deviance information criterion value.

Table 5.2. Summary of the Generalised Pareto model estimation results.

Model	Parameter	Scale			Shape	DIC
		$\emptyset_0$	$\emptyset_{PC}$	$\emptyset_{VC}$	ξ_0	
Stationary	mean	-0.064	—	—	-0.209	605
	s.d.	0.027	—	—	0.158	
	2.5%	-0.164	—	—	-0.351	
	97.5%	-0.007	—	—	-0.040	
Pedestrian volume	mean	0.283	-0.08	—	-0.274	586
	s.d.	0.087	0.062	—	0.164	
	2.5%	0.202	-0.133	—	-0.422	
	97.5%	0.373	-0.01	—	-0.1	
Vehicle volume	mean	0.379	—	-0.04	-0.292	569
	s.d.	0.380	—	0.026	0.131	
	2.5%	0.003	—	-0.064	-0.407	
	97.5%	0.748	—	-0.013	-0.15	
Pedestrian and vehicle volumes	mean	0.455	-0.056	-0.033	-0.322	565
	s.d.	0.42	0.071	0.027	0.155	
	2.5%	0.011	-0.102	-0.058	-0.457	
	97.5%	0.835	-0.001	-0.005	-0.154	

Abbreviations: PC = pedestrian volume; VC = vehicle volume; DIC = Deviance Information Criterion; s.d. = standard deviation.

Figure 5.4 shows the goodness-of-fit of the selected model, which is used to further assess the model performance using the probability density plot and quartile comparison plot of the empirical and modelled negated post encroachment time. The following observations can be made when visually inspecting the graphs below; Figure 5.4 (a) indicates reasonable proximity of observations to the line-of-equity, and Figure 5.4 (b) shows that the modelled and observed probability density curves are very close to each other.



(a) modelled versus observed quartile plot (b) probability density plot

Figure. 5.4. Best-fit Generalised Pareto model (pedestrian and vehicle volume).

For the Peak Over Threshold sampling-based models, the model with covariates relating to pedestrian and vehicle volume outperforms other competing models in terms of both local and global goodness-of-fit measures. Note that using two measures has shown to rigorously improve the model performance, see Hussain et al. (2022) for more details.

Discussion

All model covariates are statistically significant as assessed by Bayesian credible intervals and logically, providing insights into vehicle-pedestrian interactions. The number of vehicles and pedestrians used as covariates in the study denotes the total count of vehicles and pedestrians respectively present within a 100m radius circle of the conflict in each snippet/episode. Both vehicle and pedestrian counts are found to be negatively associated with crash risk. Different from some other studies in the past (where road user volume data collected is an indicator of risk exposure (Ali et al., 2023b)), both vehicle and pedestrian count captured in this study are an indicator of the overall road traffic situation. A higher vehicle count in a snippet/episode represents a lower level of service that translates to lower travel speeds and better driver attention when compared to a low vehicle count. Also, pedestrians are less likely to cross the road from non-designated pedestrian crossings in high vehicle count situations, further reducing the risk of a collision. The study concurs with the safety-in-number phenomenon, implying a larger number of pedestrians in a snippet/scenario are easier to spot for drivers making the pedestrians present in that episode less likely to be the victim of a crash. On the other hand, the average speed of vehicles in the episode is found to be positively associated with the crash risk. Vehicles travelling at faster speeds are likely to exhibit harsh braking whilst interacting with pedestrians (Ali et al., 2022b), increasing the chances of being engaged in pedestrian crashes (Haque and Washington, 2015). Song et al. (2017) investigated multiple factors influencing vehicle-pedestrian crashes and concluded that the vehicle's speed

had the highest correlation with crash risk as drivers travelling at a higher speed require a longer time to apply brakes in emergencies. The average walking speed of pedestrians is found to be negatively correlated with crash risk, implying that slow-walking pedestrians are more likely to be involved in a crash. A high pedestrian walk speed would minimise the interaction between pedestrians and other vehicles, thereby reducing the crash risk.

Using these goodness-of-fit measures for comparing the Block Maxima and Peak Over Threshold sampling-based models, this study finds that the former model generates better pedestrian crash risk estimates. More specifically, the confidence interval of the Peak Over Threshold sampling-based model is found to be about four times wider confidence interval than the Block Maxima sampling-based model. This finding is in contrast with most of the literature whereby Peak Over Threshold sampling-based models are often reported to outperform Block Maxima sampling-based models (see Hussain et al. (2022) and Zheng et al. (2014)), and this better performance is attributed to better utilisation of data. In practice, the data characteristics and the specific research objectives can determine which method is more suitable for a given analysis. Bücher and Zhou (2021) compared the two sampling approaches to better understand their differences. They summarised that the data-generating process could affect the convergence rate of the two methods with no general winner identifiable between the two approaches. They also pointed out that under independent identically distributed scenario assumption for a time series data (as is the case in this study), Block Maxima sampling works because the extremes are approximately Generalised Extreme Value-distributed and are sufficiently distant from each other to bear low serial dependence. In summary, there is no overall clear winner between the two sampling techniques, but the reasons mentioned above can provide some insights into the outcomes specific to this study.

5.2.Global model performance

The global model performance section, as described earlier in methodology section 3.4, focuses on evaluating the model's performance against real-life observed crash data. This section presents the results obtained through model simulations for the modelled crash confidence interval and compares them with the observed crash confidence interval obtained using the Poisson confidence interval statistical method as described in Eq. 3.5 in the earlier section.

Estimation results from the Block Maxima and Peak Over Threshold sampling-based models and their comparison to the observed crashes are presented in Table 5.3. To estimate crash frequency from selected models for a given duration $\tilde{T}$, Eq. 3.8 was used. The confidence

interval of estimate crashes was calculated using a simulation process. The observed crashes were sourced from Florida DOT database as described in chapter 4.4. The confidence interval for observed crashes was calculated using the Poisson confidence interval for true mean as described in Eq. 3.9. The results suggest that the Block Maxima sampling-based model outperforms the Peak Over Threshold sampling-based model when compared to observed crashes. The relative error (calculated as the estimated crashes – observed crashes) of the Block Maxima sampling-based model is 15%, whereas the corresponding error of the Peak Over Threshold sampling-based model is 90%, suggesting a six-fold higher error in the Peak Over Threshold sampling-based model. From Table 5.3, it is also evident that confidence intervals for both models are wider, which can be attributed to the relatively small dataset size within the study area. However, the confidence interval of the Block Maxima sampling-based model is relatively narrower compared to its counterpart.

Table 5.3. Estimation of crash frequencies by the developed extreme value models.

Model	Annual crashes	Confidence interval	Relative crash error (against observed)	Crash confidence interval comparison (against observed)
Block Maxima	8.1	(0, 116.1)	15%	10 times observed
Peak Over Threshold	13.3	(0, 406.0)	90%	35 times observed
Observed Crashes	7	(2.81,14.42)	-	

Pedestrians are a vulnerable road user group exposed to a significant risk of injuries and fatalities, requiring elevated research efforts to understand the determinants of pedestrian crashes and to devise targeted countermeasures. Conventionally, pedestrian safety is assessed using the reactive approach based on police-reported data, requiring unusually high (and slowly accruing) pedestrian crashes. Another issue with such data is the lack of information about crashes occurring at a corridor level. Overcoming this research need, this study presents a framework that leverages autonomous vehicle data to estimate corridor-wide pedestrian-vehicle crash risk. The underpinning of this framework is Bayesian generalised non-stationary modelling approach. Applying the framework to freely available Argoverse autonomous vehicle Level 4 data, the Block Maxima and Peak Over Threshold sampling-based models are estimated, with the mean estimated crashes of the Block Maxima sampling-based model being within a 15% error margin of the observed crashes. Due to the limited sample size, the confidence intervals of both models are wider than the observed confidence interval, whereas a comparison of the confidence intervals of both models suggests that the Block Maxima sampling-based model possesses a narrower confidence interval than its competing model. In

summary, our results confirmed the efficacy of the proposed framework for estimating vehicle-pedestrian crashes at a corridor level with reasonable accuracy.

To summarise, this study is the first application—to the best of the authors' knowledge—of utilising publicly available autonomous vehicle sensor data for assessing corridor-level pedestrian safety. A multi-sensor setup on autonomous vehicles used for data collection in this study and data integration from multiple sources can resolve most trajectory processing errors commonly seen in data-capturing practices (Wu et al., 2018). Extreme value theory models developed in this study show the capability of combining data of multiple intersection intersections and mid-block sections along the study corridor, which can improve crash estimation efficiency in a real-world application.

Chapter 6 Conclusion and recommendations

This study is one of the first to apply open-source autonomous vehicle sensor data to conduct a corridor-wide pedestrian safety analysis. The data processing framework is modular and scalable to larger problem sets. The study developed traffic conflict-based univariate extreme value models for vehicle-pedestrian safety analysis along an urban corridor. Best-fitted models were then compared against observed crash data from the same corridor. The study successfully demonstrated that a robust data processing and filtering framework could extract useful trajectory information from autonomous vehicle sensor data. The study successfully developed a combined safety model for intersections and mid-block sections which have typically been dealt with separate models in past research (Kamel et al., 2022). As autonomous vehicles become mainstream and their adoption rate increase, autonomous vehicle sensor data has an inherent potential to deal with data sparsity issues from traditional video analytics data sources.

This study answered the following research questions through the proposed framework and its implementation:

- How to process and interpret autonomous vehicle sensor data in a meaningful way?

  The raw autonomous vehicle data extracted from Argoverse, including over 250,000 episodes, was organised, indexed, and categorised to facilitate its usage. The dataset was then converted to a tabular data frame structure for efficient processing. Object trajectories were constructed from all the episodes to capture the movement of various road users, and the coordinate reference system was transformed to allow data integration from multiple sources. This process generated a standardised data output from the process that has universal interpretability and does not require any special tools/software for usage.

- How to extract pedestrian-vehicle conflict information from autonomous vehicle sensor data and check the quality of the output?

  The dataset was split into two data frames containing pedestrian trajectories and vehicle trajectories, respectively. Overlaying trajectories of each possible pair of objects from two data frames, the potential conflicts were identified. Subsequently, an encroachment zone was defined for the potential conflicting trajectories to extract conflict indicator information. In order to ensure the quality of trajectory data, trajectory sanity checks

involving cross-referencing geographical and contextual information from map data and scenario data were performed.

- How to apply Extreme Value Theory modelling technique for a corridor-wide safety assessment framework?

  An Extreme Value Theory model was developed using 581 unique traffic conflict observations collected using autonomous vehicle data along a 19-kilometre-long study corridor. Fitted models estimated using two extreme value sampling techniques, block maxima and peak-over-threshold, were assessed using deviance information criteria and model validation against observed crashes. Various modelling covariates, such as pedestrian count, vehicle count, average pedestrian speed, and average vehicle speed, were introduced to address the model's unobserved heterogeneity. The results demonstrate that the models generated a reasonable fit against the observed crash data along the corridor. The best-performing model estimated annual crashes within 15% of the observed crashes, demonstrating the applicability of this technique for a corridor-level analysis.

The future scope of improvement to work done in this study includes testing safety metrics incorporating a more comprehensive array of factors influencing crash risk across a corridor to improve focus on key problem areas on a network. This work determined that the data annotation algorithm used by data providers still has room for improvement. Research on handling mislabelled data more efficiently would improve the real-time application prospects of such models. More research is needed to understand the difference in the interaction between autonomous vehicle-autonomous vehicle and autonomous vehicle-other road users. As the penetration rate in this dataset was shallow, no autonomous vehicle-autonomous vehicle interaction samples were available.

Furthermore, research should be undertaken to understand the impact of an increase in quantity and change in autonomous vehicle sensor data format as the dataset is still primitive, only providing discrete episode outputs instead of a continuous data stream. In the future, using autonomous vehicle sensor data supplemented with additional data sources can aid in developing a network-wide safety model incorporating multiple road user interactions. Furthermore, future studies can incorporate other corridor-based variables such as lane attributes, traffic manoeuvre counts. The crash severity aspect can be included into the model structure by employing multivariate model structure with conflict indicator dedicated to crash

severity. This extension of the model will enable a more comprehensive understanding of the potential factors influencing the severity of the crashes, which can improve the effectiveness and focus of road safety interventions.

Reference

Ali, Y., Haque, M. & Zheng, Z. 2022a. Assessing a Connected Environment's Safety Impact During Mandatory Lane-Changing: A Block Maxima Approach. *IEEE Transactions on Intelligent Transportation Systems*, 1–11. https://doi.org/10.1109/TITS.2022.3147668

Ali, Y., Haque, M. M. & Mannering, F. 2023a. Assessing traffic conflict/crash relationships with extreme value theory: Recent developments and future directions for connected and autonomous vehicle and highway safety research. *Analytic Methods in Accident Research,* 39. https://doi.org/10.1016/j.amar.2023.100276

Ali, Y., Haque, M. M. & Mannering, F. 2023b. A Bayesian generalised extreme value model to estimate real-time pedestrian crash risks at signalised intersections using artificial intelligence-based video analytics. *Analytic Methods in Accident Research,* 38. https://doi.org/10.1016/j.amar.2022.100264

Ali, Y., Haque, M. M., Zheng, Z. & Afghari, A. P. 2022b. A Bayesian correlated grouped random parameters duration model with heterogeneity in the means for understanding braking behaviour in a connected environment. *Analytic Methods in Accident Research,* 35. https://doi.org/10.1016/j.amar.2022.100221

Alozi, A. R. & Hussein, M. 2022. Evaluating the safety of autonomous vehicle–pedestrian interactions: An extreme value theory approach. *Analytic Methods in Accident Research,* 35. https://doi.org/10.1016/j.amar.2022.100230

Ammar, D., Xu, Y., Jia, B. & Bao, S. 2022. Examination of Recent Pedestrian Safety Patterns at Intersections through Crash Data Analysis. *Transportation Research Record,* 2676, 331–341. https://doi.org/10.1177/03611981221095513

Arun, A., Haque, M. M., Bhaskar, A. & Washington, S. 2022. Transferability of multivariate extreme value models for safety assessment by applying artificial intelligence-based video analytics. *Accident Analysis and Prevention,* 170. https://doi.org/10.1016/j.aap.2022.106644

Arun, A., Haque, M. M., Bhaskar, A., Washington, S. & Sayed, T. 2021a. A bivariate extreme value model for estimating crash frequency by severity using traffic conflicts. *Analytic Methods in Accident Research,* 32. https://doi.org/10.1016/j.amar.2021.100180

Arun, A., Haque, M. M., Bhaskar, A., Washington, S. & Sayed, T. 2021b. A systematic mapping review of surrogate safety assessment using traffic conflict techniques. *Accident Analysis and Prevention,* 153, 106016–106016. https://doi.org/10.1016/j.aap.2021.106016

Arun, A., Haque, M. M., Washington, S., Sayed, T. & Mannering, F. 2021c. A systematic review of traffic conflict-based safety measures with a focus on application context. *Analytic Methods in Accident Research,* 32. https://doi.org/10.1016/j.amar.2021.100185

Arun, A., Lyon, C., Sayed, T., Washington, S., Loewenherz, F., Akers, D., Ananthanarayanan, G., Shu, Y., Bandy, M. & Haque, M. M. 2023. Leading pedestrian intervals – Yay or Nay? A Before-After evaluation of multiple conflict types using an enhanced Non-Stationary framework integrating quantile regression into Bayesian hierarchical extreme value analysis. *Accident Analysis and Prevention,* 181, 106929–106929. https://doi.org/10.1016/j.aap.2022.106929

Beauchamp, É., Saunier, N. & Cloutier, M.-S. 2022. Study of automated shuttle interactions in city traffic using surrogate measures of safety. *Transportation Research. Part C, Emerging Technologies,* 135. https://doi.org/10.1016/j.trc.2021.103465

Bendak, S., Alnaqbi, A. M., Alzarooni, M. Y., Aljanaahi, S. M. & Alsuwaidi, S. J. 2021. Factors affecting pedestrian behaviors at signalized crosswalks: An empirical study. *Journal of Safety Research,* 76, 269-275. https://doi.org/10.1016/j.jsr.2020.12.019

Bernhardt, M. & Kockelman, K. 2021. An analysis of pedestrian crash trends and contributing factors in Texas. *Journal of Transport & Health,* 22. https://doi.org/10.1016/j.jth.2021.101090

Bücher, A. & Zhou, C. 2021. A horse racing between the block maxima method and the peak-over-threshold approach. *Statistical Science,* 36**,** 360–378. https://doi.org/10.1214/20-STS795

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G. & Beijbom, O. nuScenes: A multimodal dataset for autonomous driving. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, 2019. 11621-11631.

Char, F. & Serre, T. 2020. Analysis of pre-crash characteristics of passenger car to cyclist accidents for the development of advanced drivers assistance systems. *Accident Analysis and Prevention,* 136**,** 105408–105408. https://doi.org/10.1016/j.aap.2019.105408

Coles, S. 2001. *An Introduction to Statistical Modeling of Extreme Values,* London, Springer. https://doi.org/10.1007/978-1-4471-3675-0

Essa, M. & Sayed, T. 2019. Full Bayesian conflict-based models for real time safety evaluation of signalized intersections. *Accident Analysis and Prevention,* 129**,** 367–381. https://doi.org/10.1016/j.aap.2018.09.017

Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J. & Anguelov, D. Large scale interactive motion forecasting for autonomous driving : The waymo open motion dataset. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 9690–9699. https://doi.org/10.1109/ICCV48922.2021.00957

FLHSMV, F. H. S. a. M. V. 2022. *TRAFFIC CRASH REPORTS Crash Dashboard* [Online]. Available: https://www.flhsmv.gov/traffic-crash-reports/crash-dashboard/ [Accessed 03/09/2022].

Fu, C. & Sayed, T. 2021. Multivariate Bayesian hierarchical Gaussian copula modeling of the non-stationary traffic conflict extremes for crash estimation. *Analytic Methods in Accident Research,* 29. https://doi.org/10.1016/j.amar.2020.100154

Geiger, A., Lenz, P., Stiller, C. & Urtasun, R. 2013. Vision meets robotics: The KITTI dataset. *The International Journal of Robotics Research,* 32**,** 1231–1237. https://doi.org/10.1177/0278364913491297

Ghadirzadeh, S., Mirbaha, B. & Rassafi, A. A. 2022. Analysing pedestrian–vehicle conflict behaviours at urban pedestrian crossings. *Proceedings of the Institution of Civil Engineers. Municipal Engineer,* 175**,** 107–118. https://doi.org/10.1680/jmuen.21.00016

Gooch, J., Hamilton, I., Polin, B., Tanzen, R. & Cohen, T. 2022. Systemic Safety Analysis of Midblock Pedestrian Crashes in Massachusetts. *Transportation Research Record,* 2676**,** 722–730. https://doi.org/10.1177/03611981221094566

Guido, G., Saccomanno, F., Vitale, A., Astarita, V. & Festa, D. 2011. Comparing Safety Performance Measures Obtained from Video Capture Data. *Journal of Transportation Engineering, Part A,* 137**,** 481–491. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000230

Guo, F., Klauer, S. G., Hankey, J. M. & Dingus, T. A. 2010. Near Crashes as Crash Surrogate for Naturalistic Driving Studies. *Transportation Research Record,* 2147**,** 66–74. https://doi.org/10.3141/2147-09

Guo, Y., Sayed, T. & Zheng, L. 2020. A hierarchical bayesian peak over threshold approach for conflict-based before-after safety evaluation of leading pedestrian intervals.

Accident Analysis and Prevention, 147**,** 105772–105772. https://doi.org/10.1016/j.aap.2020.105772

Haque, M. M. & Washington, S. 2015. The impact of mobile phone distraction on the braking behaviour of young drivers: A hazard-based duration model. *Transportation Research. Part C, Emerging Technologies*, 50**,** 13–27. https://doi.org/10.1016/j.trc.2014.07.011

Hong, J., Shankar, V. N. & Venkataraman, N. 2016. A spatially autoregressive and heteroskedastic space-time pedestrian exposure modeling framework with spatial lags and endogenous network topologies. *Analytic Methods in Accident Research*, 10**,** 26–46. https://doi.org/10.1016/j.amar.2016.05.001

Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., Omari, S., Iglovikov, V. & Ondruska, P. 2020. One Thousand and One Hours: Self-driving Motion Prediction Dataset. *arXiv*, 2006. https://doi.org/10.48550/arxiv.2006.14480

Hu, X., Zheng, Z., Chen, D., Zhang, X. & Sun, J. 2022. Processing, assessing, and enhancing the Waymo autonomous vehicle open dataset for driving behavior research. *Transportation Research. Part C, Emerging Technologies*, 134. https://doi.org/10.1016/j.trc.2021.103490

Hussain, F., Li, Y., Arun, A. & Haque, M. M. 2022. A hybrid modelling framework of machine learning and extreme value theory for crash risk estimation using traffic conflicts. *Analytic Methods in Accident Research*, 36. https://doi.org/10.1016/j.amar.2022.100248

Hydén, C. 1987. The development of a method for traffic safety evaluation: the Swedish traffic conflicts technique. Lund, Sweden: Lund Institute of Technology.

IHME, I. f. H. M. a. E. 2021. Global Burden of Disease Study 2019. *In:* NETWORK, G. B. O. D. C. (ed.).

Ismail, K., Sayed, T. & Saunier, N. 2011. Methodologies for Aggregating Indicators of Traffic Conflict. *Transportation Research Record*, 2237**,** 10–19. https://doi.org/10.3141/2237-02

Job, R. F. S. 2020. Policies and Interventions to Provide Safety for Pedestrians and Overcome the Systematic Biases Underlying the Failures. *Frontiers in Sustainable Cities*, 2. https://doi.org/10.3389/frsc.2020.00030

Kamel, A., Sayed, T. & Fu, C. 2022. Real-time safety analysis using autonomous vehicle data: a Bayesian hierarchical extreme value model. *Transportmetrica***,** 1-21. https://doi.org/10.1080/21680566.2022.2135634

Kaparias, I., Bell, M. G. H., Greensted, J., Cheng, S., Miri, A., Taylor, C. & Mount, B. 2010. Development and Implementation of a Vehicle–Pedestrian Conflict Analysis Method: Adaptation of a Vehicle–Vehicle Technique. *Transportation Research Record*, 2198**,** 75–82. https://doi.org/10.3141/2198-09

Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A. & Ondruska, P. 2019. *Lyft level 5 av dataset 2019* [Online]. Available: https://level5.lyft.com/dataset [Accessed 10/05/2020].

Khayesi, M. 2020. Vulnerable Road Users or Vulnerable Transport Planning? *Frontiers in Sustainable Cities*, 2. https://doi.org/10.3389/frsc.2020.00025

Kutela, B., Das, S. & Dadashova, B. 2022. Mining patterns of autonomous vehicle crashes involving vulnerable road users to understand the associated factors. *Accident Analysis and Prevention*, 165**,** 106473–106473. https://doi.org/10.1016/j.aap.2021.106473

Laureshyn, A., Svensson, Å. & Hydén, C. 2010. Evaluation of traffic safety, based on micro-level behavioural data: Theoretical framework and first implementation. *Accident Analysis and Prevention*, 42**,** 1637–1646. https://doi.org/10.1016/j.aap.2010.03.021

Li, C., Liu, S. & Cen, X. 2021. Safety and efficiency impact of pedestrian–vehicle conflicts at non signalized midblock crosswalks based on fuzzy cellular automata. *Physica A*, 572. https://doi.org/10.1016/j.physa.2021.125871

Li, S., Cai, B., Liu, J. & Wang, J. 2018. Collision risk analysis based train collision early warning strategy. *Accident Analysis and Prevention*, 112**,** 94–104. https://doi.org/10.1016/j.aap.2017.11.039

Liu, Q., Wang, X., Wu, X., Glaser, Y. & He, L. 2021. Crash comparison of autonomous and conventional vehicles using pre-crash scenario typology. *Accident Analysis and Prevention*, 159**,** 106281–106281. https://doi.org/10.1016/j.aap.2021.106281

Ma, Y., Qin, X., Grembek, O. & Chen, Z. 2018. Developing a safety heatmap of uncontrolled intersections using both conflict probability and severity. *Accident Analysis and Prevention*, 113**,** 303–316. https://doi.org/10.1016/j.aap.2018.01.038

Pawar, D. S. & Patil, G. R. 2017. Minor-Street Vehicle Dilemma While Maneuvering at Unsignalized Intersections. *Journal of Transportation Engineering, Part A*, 143. https://doi.org/10.1061/JTEPBS.0000066

Rahman, M. S., Abdel-Aty, M., Lee, J. & Rahman, M. H. 2019. Safety benefits of arterials' crash risk under connected and automated vehicles. *Transportation Research. Part C, Emerging Technologies*, 100**,** 354–371. https://doi.org/10.1016/j.trc.2019.01.029

Santhosh, D., Bindhu, B. K. & Koshy, B. I. 2020. Evaluation of pedestrian safety in unsignalized T and X – Intersections through comparison of the frequency and severity of pedestrian conflicts. *Case Studies on Transport Policy*, 8**,** 1352–1359. https://doi.org/10.1016/j.cstp.2020.09.006

Sharma, A., Ali, Y., Saifuzzaman, M., Zheng, Z. & Haque, M. M. 2018. Human Factors in Modelling Mixed Traffic of Traditional, Connected, and Automated Vehicles. *Advances in Intelligent Systems and Computing*, 591. https://doi.org/10.1007/978-3-319-60591-3_24

Smith, C. L. 2020. Representing external hazard initiating events using a Bayesian approach and a generalized extreme value model. *Reliability Engineering & System Safety*, 193. https://doi.org/10.1016/j.ress.2019.106650

Sobhani, A., Young, W., Bahrololoom, S. & Sarvi, M. 2013. Calculating time-to-collision for analysing right turning behaviour at signalised intersections. *Road & Transport Research*, 22**,** 49–61.

Song, T.-J., So, J., Lee, J. & Williams, B. M. 2017. Exploring Vehicle–Pedestrian Crash Severity Factors on the Basis of In-Car Black Box Recording Data. *Transportation Research Record*, 2659**,** 148–154. https://doi.org/10.3141/2659-16

Songchitruksa, P. & Tarko, A. P. 2006. The extreme value theory approach to safety estimation. *Accident Analysis and Prevention*, 38**,** 811-822. https://doi.org/10.1016/j.aap.2006.02.003

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhao, S., Cheng, S. & Zhang, Y. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020), 2020. 2446-2454.

Sun, X., Lin, K., Wang, Y., Ma, S. & Lu, H. 2022. A Study on Pedestrian–Vehicle Conflict at Unsignalized Crosswalks Based on Game Theory. *Sustainability (Basel, Switzerland)*, 14. https://doi.org/10.3390/su14137652

Tageldin, A., Sayed, T. & Wang, X. 2015. Can Time Proximity Measures be Used as Safety Indicators in All Driving Cultures?: Case Study of Motorcycle Safety in China. *Transportation Research Record*, 2520**,** 165–174. https://doi.org/10.3141/2520-19

Tarko, A. P. 2012. Use of crash surrogates and exceedance statistics to estimate road safety. *Accident Analysis and Prevention,* 45**,** 230–240. https://doi.org/10.1016/j.aap.2011.07.008

TMR, T. a. M. R. 2022. Road crash locations. *In:* GOVERNMENT, Q. (ed.). Open Data Portal: Queensland Government.

Wang, C., Xu, C., Xia, J., Qian, Z. & Lu, L. 2018. A combined use of microscopic traffic simulation and extreme value methods for traffic safety evaluation. *Transportation Research. Part C, Emerging Technologies,* 90**,** 281–291. https://doi.org/10.1016/j.trc.2018.03.011

Weng, J. & Meng, Q. 2014. Rear-end crash potential estimation in the work zone merging areas: rear-end crash potential estimation. *Journal of Advanced Transportation,* 48**,** 238–249. https://doi.org/10.1002/atr.211

WHO, W. H. O. 2022. *Road traffic injuries* [Online]. World Health Organization. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries [Accessed 02/01/2023 2023].

WHO, W. H. O. 2023. THE GLOBAL HEALTH OBSERVATORY. *In:* ORGANIZATION, W. H. (ed.). azureedge.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., Ramanan, D., Carr, P. & Hays, J. 2021. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021).*

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., Ramanan, D., Carr, P. & Hays, J. 2023. Argoverse 2: Next Generation Datasets for Self-Driving Perception and Forecasting. *arXiv,* 2301. https://doi.org/10.48550/arxiv.2301.00493

Wu, J. & Xu, H. 2018. The influence of road familiarity on distracted driving activities and driving operation using naturalistic driving study data. *Transportation Research. Part F, Traffic Psychology and Behaviour,,* 52**,** 75-85. https://doi.org/10.1016/j.trf.2017.11.018

Wu, J., Xu, H., Zheng, Y. & Tian, Z. 2018. A novel method of vehicle-pedestrian near-crash identification with roadside LiDAR data. *Accident Analysis and Prevention,* 121**,** 238–249. https://doi.org/10.1016/j.aap.2018.09.001

Zhang, L., Wang, L., Zhou, K. & Zhang, W.-B. 2012. Dynamic All-Red Extension at a Signalized Intersection: A Framework of Probabilistic Modeling and Performance Evaluation. *IEEE Transactions on Intelligent Transportation Systems,* 13**,** 166–179. https://doi.org/10.1109/TITS.2011.2166070

Zhao, X., Wang, Z., Xu, Z., Wang, Y., Li, X. & Qu, X. 2020. Field experiments on longitudinal characteristics of human driver behavior following an autonomous vehicle. *Transportation Research. Part C, Emerging Technologies,* 114**,** 205–224. https://doi.org/10.1016/j.trc.2020.02.018

Zheng, L., Ismail, K. & Meng, X. 2014. Freeway safety estimation using extreme value theory approaches: A comparative study. *Accident Analysis and Prevention,* 62**,** 32–41. https://doi.org/10.1016/j.aap.2013.09.006

Zheng, L., Ismail, K., Sayed, T. & Fatema, T. 2018. Bivariate extreme value modeling for road safety estimation. *Accident Analysis and Prevention,* 120**,** 83–91. https://doi.org/10.1016/j.aap.2018.08.004

Zheng, L. & Sayed, T. 2020. A novel approach for real time crash prediction at signalized intersections. *Transportation Research. Part C, Emerging Technologies,* 117. https://doi.org/10.1016/j.trc.2020.102683

Zheng, L., Sayed, T. & Mannering, F. 2021. Modeling traffic conflicts for use in road safety analysis: A review of analytic methods and future directions. *Analytic Methods in Accident Research*, 29**,** 100142. https://doi.org/10.1016/j.amar.2020.100142